Image Fidelity Assessment and its Applications

Beeldgetrouwheidsbeoordeling en haar toepassingen

Benhur Ortiz Jaramillo

Promotoren: prof. dr. ir. W. Philips, dr. L. Platisa
Proefschrift ingediend tot het behalen van de graad van
Doctor in de ingenieurswetenschappen

UNIVERSITEIT
GENT

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2017 - 2018

ISBN 978-94-6355-149-6
NUR 958
Wettelijk depot: D/2018/10.500/67

Members of the jury

Prof. Dr. Ir. Wilfried Philips (Professor at Universiteit Gent, supervisor)
Dr. Ir. Ljiljana Platisa (Postdoctoral researcher at Universiteit Gent, co-supervisor)
Prof. Dr. Ir. Aleksandra Pizurica (Professor at Universiteit Gent)
Dr. Ir. Jan Aelterman (Postdoctoral researcher at Universiteit Gent)
Dr. Ir. Glenn Van Wallendael (Postdoctoral researcher at Universiteit Gent)
Prof. Dr. Ir. Peter Schelkens (Professor at Vrije Universiteit Brussel)
Prof. Dr. Ir. Patrick Le Callet (Professor at Ecole polytechnique de l'université de Nantes)

Affiliations

Research Group for Image Processing and Interpretation (IPI)
Independent research center imec
Department of Telecommunications and Information Processing (TELIN)
Faculty of Engineering and Architecture
Ghent University

Sint-Pietersnieuwstraat 41
B-9000 Ghent
Belgium

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur
Vakgroep Telecommunicatie en Informatieverwerking

Promotoren: Prof. Dr. Ir. Wilfried Philips
Dr. Ir. Ljiljana Platisa

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur

Vakgroep Telecommunicatie en Informatieverwerking
Sint-Pietersnieuwstraat 41, B-9000 Gent, België

Tel.: +32-9-264.79.66
Fax.: +32-9-264.42.95

Voorzitter: Prof. Dr. Ir. Herwig Bruneel

Proefschrift ingediend tot het behalen van de graad van
Doctor in de ingenieurswetenschappen
Academiejaar 2017 - 2018

Acknowledgements

This dissertation would never have been accomplished without the guidance, help and support of many people. Especially I would like to express my gratitude to the following people: I would like to thank my advisor, Prof. Dr. Ir. Wilfried Philips, for giving me the opportunity to conduct my doctoral research at Image Processing and Interpretation (IPI) group, Ghent University and for sharing his immense knowledge in the image processing field as well as for his constant support. There are not enough words to describe my gratitude to my co-advisor Dr. Ir. Ljiljana Platisa. Without her supervision, support, advice, guidance, patience, motivation, constructive feedback, enthusiasm and immense knowledge in image fidelity assessment this dissertation would not be possible.

I am grateful to my (ex)colleagues from the Department of Telecommunications and Information Processing (TELIN) specially those who have shared research or personal moments with me: Prof. Dr. Ir. Herwig Bruneel, Asli Kumcu, Danilo Babin, Angel Lopez, Ljubomir Jovanov, Jose Menendez, Patrick Schaillée, Wenzhi Liao, Gonzalo Luzardo, Jorge Rodriguez, Dimitri Van Cauwelaert, Sergio Orjuela, Renbo Luo and Mario Pinto.

Special thanks to the members of the jury for taking the time to revise this dissertation: Prof. Dr. Ir. Aleksandra Pizurica, Dr. Ir. Jan Aelterman, Dr. Ir. Glenn Van Wallendael, Prof. Dr. Ir. Peter Schelkens and Prof. Dr. Ir. Patrick Le Callet.

I would like to thank my dear friends: Rolando Quinoñes and his wonderful family (Rocio and Nicolas), Nuno Arroz, Alfredo Matia, Andrea Teufelberger, Sergio Candela, Tanja Consolati, Jennifer Triana, Jose Armando Fernandez, Carlos Herrera, Johana Sandoval, Mercedes Caron and Francesco Iannaccone. Special thanks to my dear friend Charlotte Lievens. Also, I want to thank the whole Kafka's futsal team who provided me a healthy environment to relax between the research periods.

The document of this dissertation would not have been ended successfully without the emotional support of my mother Haydee Jaramillo, my two brothers Carlos and Enrique Ortiz, and my beloved girlfriend Alexandrina Roca.

Ghent, February, 2018.
Benhur Ortiz-Jaramillo

Samenvatting

Het aantal toepassingen dat vertrouwt op digitale beeldvorming blijft jaar na jaar toenemen. Bijvoorbeeld, het waarneembare licht is verworven voor digitale fotografie; röntgenstralen laten het gebruik van digitale beeldvorming toe voor medische toepassingen, zoals fluoroscopie. De meeste toepassingen resulteren in beelden en video's bestemd om te worden bekeken door mensen. Daarom is de beoordeling van beeldgetrouwheid belangrijk: het objectief beoordelen van waargenomen verschillen tussen een referentiebeeld, en één of meer overeenkomstige testbeelden.

Historisch gezien worden de termen beeldgetrouwheid en beeldkwaliteit door elkaar gebruikt, maar ze betekenen niet helemaal hetzelfde: De beoordeling van beeldgetrouwheid verwijst naar het meten van waarneembare verschillen tussen twee beelden: een referentiebeeld en een testbeeld. De beoordeling van beeldkwaliteit daarentegen verwijst naar het beoordelen door middel van subjectieve voorkeur voor een beeld. Bijvoorbeeld: wanneer men in het kader van beeldverbetering twee testbeelden naast mekaar houdt, is het mogelijk dat menselijke waarnemers in een van deze twee beelden minder verschillen met het referentiebeeld detecteren, en desondanks het testbeeld verkiezen met de meeste verschillen.

Tegenwoordig probeert de meerderheid van de state-of-the-art methodes beeldgetrouwheid te voorspellen volgens "one-size-fits-all"- oplossingen, gebaseerd op de laatste inzichten op vlak van het menselijk visueel systeem. Die aanpak resulteert doorgaans in ingewikkelde computeralgoritmes. Bijgevolg zijn deze algoritmes niet wenselijk voor gebruik in real-time beeldverwerkingsalgoritmes of systemen. Daarbovenop is het voor vele toepassingen net wenselijk gebruik te maken van toepassings-specifieke perceptuele kenmerken. Wanneer we bijvoorbeeld verschillen in het uitzicht van textielweefsels bepalen, is het van bijzonder belang de structuurkenmerken te vergelijken van de te analyseren oppervlakken.

In deze thesis onderzoeken we getrouwheidsbeoordeling voor verschillende toepassingen zoals videocompressie, contrastwijziging van digitale beelden, evaluatie van uiterlijke veranderingen van textuur, en het vaststellen van kleurverschillen bij natuurlijke beelden.

Allereerst bestuderen we de meest gekende taak in getrouwheidsbeoordeling: vaststellen in welke mate een gecomprimeerde video of beeld beantwoordt aan een bepaalde referentie (een "perfecte" video, vrij van ruis); met andere woorden: de objectieve kwalitatieve evaluatie van gecomprimeerde videosequenties. Deze thesis stelt een methodologie voor om bestaande kwaliteitsmeetmeth-

odes voor video te verbeteren, door videoinhoudsgerelateerde indexen aan hun berekeningswijze toe te voegen. De voorgestelde methode is minder complex dan conventionele methodes, en komt zelfs tegemoet aan de vereisten van real-time toepassingen. Toch is de nauwkeurigheid ervan vergelijkbaar met die van conventionele methodes. Daarmee hebben we Python-software ontwikkeld die in staat is waargenomen videokwaliteit te berekenen bij 12, 25 en 75 frames per seconde voor respectivelijk 1920×1080, 1280×720 en 720×380 pixels.

Ten tweede onderzoekt deze thesis de evaluatie van contrastverhoudingen in beelden. We bestuderen in het bijzonder de evaluatie van veranderingen in contrastverhouding tussen twee beelden: een referentiebeeld (het standaard staal) en een testbeeld (het beeld na aanpassingen in contrast). We stellen een nieuwe methode voor om contrastverhoudingen van beelden te berekenen door stukken beeld te kenmerken door middel van hun bimodale histogrammen. We gebruiken Weber en Michelson-formules voor contrastverhouding op de vlakken, om een situatie te simuleren waarbij respectievelijk een kleine, interessante structuur aanwezig is op een uniforme achtergrond, of een square-wave grating van een cyclus. De voorgestelde methode voorspelt accuraat, en beter dan andere state-of-the-art algoritmes, verschillen in beeld ten gevolge van contrastverhoudingen (afnames en toenames). We testen onze methodologie in een realistisch scenario waarbij contrastwijziging op interventionele röntgenfoto's verkregen ten gevolge van variërende stralingsdosis dient gedetecteerd te worden. De voorgestelde meting voor beeldcontrastverhouding blijkt goed overeen te stemmen met de subjectieve evaluatie door experts in de interventieradiologie.

Als derde geeft deze thesis een overzicht van methodes om veranderingen in het uitzicht van texturen te evalueren. We bespreken en evalueren veertien descriptoren voor textuuranalyse die worden gebruikt voor automatische evaluatie van verandering in het voorkomen van textiel. De besproken technieken evalueren we bij automatische evaluatie van degradatie aan het oppervlak van vloerbekledingen. We bestuderen vier categorieën van textuurbeschrijving: statistische kenmerken, structurele kenmerken, kenmerken gebaseerd op signaalverwerking en modelgebaseerde kenmerken. Ook is de impact van de parameterselectie voor de geëvalueerde textuuranalytische descriptoren bestudeerd. Hieruit blijkt dat de methodes gebaseerd op signaalverwerking de beste uitvoering geven. Deze vertonen een sterke correlatie met de expertenbeoordeling voor twee standaardtypes van oppervlakteconstructie.

Ten vierde bestuderen we methodes om waargenomen kleurverschillen te evalueren bij beelden van natuurlijke omgevingen. We evalueren achttien algoritmes voor kleurverschillen, ontworpen voor beelden van natuurlijke omgevingen. Ook stellen we een volledig nieuwe methode voor om kleurverschillen te bepalen op stukken met plaatselijk homogene textuur, dit zijn reeksen van geconnecteerde pixels met hetzelfde textuurpatroon. De basis van onze meettechniek ligt in het gegeven dat mensen kleurverschillen beoordelen door het vergelijken van reeksen geconnecteerde pixels, of van vlakken die doorgaans als homogeen aanzien worden. Bijgevolg gebruiken we beeldsegmentatie gebaseerd

op textuur, om de kleurverschillen in de resulterende segmenten te berekenen. We testen de algoritmes voor kleurverschilbepaling op drie vervormde beelden met kleurverwantschap: kwantisatieruis, gemiddelde verschuiving (verschuiving intensiteit), en verandering in kleurverzadiging. De onderzoeksresultaten tonen aan dat de voorgestelde methode beter en accurater is in het voorspellen van waargenomen veranderingen in kleur (kleurverschillen) dan de state-of-the-art algoritmes.

Bovendien hebben we gedurende deze thesis een softwaretool ontwikkeld om getrouwheidsbeoordelings meettechnieken van beelden te berekenen, ter assistentie van beeldgetrouwheidsonderzoekers. De tool biedt makkelijke toegang tot een brede waaier aan state-of-the-art meettechnieken: achttien meettechnieken voor kleurverschillen in digitale beelden, veertien algoritmes voor textuuranalyse, zes beeldcontrastverhoudingstechnieken en zes beeldkwaliteitsmeettechnieken (piek signaal-ruisverhouding, structurele gelijkenis index, blocking maatstaf, schatting ruis, blurring maatstaf en edge piek signaal-ruisverhouding). Deze state-of-the-art technieken kunnen toegepast worden op een enkel paar beelden en/of op een volledige database. Als extra ondersteuning voor data analyse laat de tool ook inuÃŕtieve visualisatie toe, zoals spreidingsdiagrammen en de resultaten van de correlatieanalyse.

Het werk dat in deze thesis werd ontwikkeld, is voorgesteld en besproken op zes internationale conferenties, in vier peer-gereviewde artikels en op een internationaal forum.

Summary

The number of applications that rely on digital imaging as means of representing information continues to increase over the years. For instance, the visible light is acquired for digital photography; x-rays allows the use of digital imaging for medical applications (e.g., fluoroscopy). Since images and videos are typically intended to be viewed by humans, a considerable attention has been paid to image fidelity assessment: the objective assessment of the perceived differences between a reference image and one or more corresponding test images.

Historically, the terms fidelity and quality have been used interchangeable in the image and video processing field, but they are often not the same. On the one hand, image fidelity assessment refers to quantifying perceptual differences between two images: a reference and test image. On the other hand, image quality assessment refers to assessing the subjective preference for one image over another. For instance, in image enhancement when comparing two test images, human observers may detect smaller differences in one of the two test images compared to the reference image and still prefer the test image with the highest difference.

Nowadays, the majority of image fidelity models in the state-of-the-art try to predict image fidelity using one-size-fits-all solutions based on the last advances in human visual system models which in general results in complex computer algorithms. Thus, those fidelity measures are cumbersome for inclusion in any real-time image processing algorithm or system. Alternatively, it is more convenient to take advantage of the individual characteristics of the perceptual differences depending on the application at hand, e.g., when measuring appearance retention of textiles, it is of particular interest to compare the texture features of the textile surfaces under analysis.

In this thesis we investigate application-tailored fidelity assessment tasks such as, quality estimation of compressed video sequences, evaluation of contrast changes in digital images, evaluation of appearance changes in texture, and color difference assessment of natural scene color images.

First, we study the most well-known fidelity assessment task: to determine how close a compressed image/video is to a given reference (distortion free or "perfect" video), i.e., the objective estimation of quality of compressed video sequences. This thesis proposes a methodology to advance existing video quality measures by introducing video content related indexes in their computation. The complexity of the proposed method is low compared to other conventional methods and it satisfies the requirements of real-time applications. At the same

time, the accuracy of the proposed method is comparable with the conventional methods. Additionally, we have implemented a Python script able to compute perceived video quality at 12, 25 and 75 frames per second for 1920×1080, 1280×720 and 720×380 pixels, respectively.

Second, this thesis investigates the evaluation of contrast ratio in images. Particularly, we study the assessment of contrast ratio changes between two images: a reference image (standard sample) and a test image (after contrast adjustment). We propose a novel methodology to compute contrast ratio in images by characterizing image patches through bimodal histograms. We use Weber and Michelson contrast ratio formulas on the patches to simulate the cases where a small structure of interest is present on a uniform background or a square-wave grating of one cycle, respectively. The proposed measure accurately predicts image differences due to contrast changes (decrements and increments) better than other state-of-the-art algorithms. We test our methodology for a real case scenario of the detection of contrast changes in interventional x-ray images acquired with varying dose. The results show that the proposed image contrast ratio measure agrees with the subjective evaluation of expert x-ray interventionalists.

Third, this dissertation surveys the methods for evaluation of appearance changes in texture. We review and evaluate fourteen texture analysis descriptors for the automatic evaluation of appearance changes in textiles. To evaluate the reviewed techniques, we consider the degradation appearing on the surface of textile floor coverings. We have studied the four texture descriptor categories: statistical features, structural features, signal processing based features and model based features. Also, the impact of the parameter selection for the evaluated texture analysis descriptors is discussed. The experimental results show that the signal processing based methods are the best performing methods, achieving a strong correlation with the textile specialists' assessment in two standard surface type constructions.

Fourth, this thesis studies the methods for evaluation of perceived color differences in natural scene color images. We evaluate eighteen color difference algorithms designed for natural scene color images. Also, we propose a novel method to compute color differences on local homogeneous textured patches, i.e., set of connected pixels with the same texture pattern. We base our measure on the fact that humans assess color differences in natural scene color images by comparing sets of connected pixels or small patches typically characterized for being homogeneous. Therefore, we use image segmentation based on texture to compute the color differences in the resulting segments. We test the color difference algorithms on three color related distortions from one publicly available database: quantization noise, mean shift (intensity shift), and change of color saturation. The results show that the proposed method is able to accurately predict perceived changes of color (color differences) better than the state-of-the-art algorithms.

Additionally, during this thesis we have developed a software tool to compute fidelity assessment measures in images designed to assist image fidelity

researchers. The developed tool provides easy access to a range of state-of-the-art measures: eighteen color difference measures for digital images, fourteen texture analysis algorithms, six image contrast ratio measures and six image quality measures (peak signal to noise ratio, structural similarity index measure, blocking measure, noise estimation, blurring measurement and edge peak signal to noise ratio). The state-of-the-art measures can be applied on a single pair of images and/or in a full database, as well as enables intuitive visualizations that aid data analysis, e.g., scatter plots and the results of the correlation analysis.

The work developed in this thesis has been presented and discussed in six international conferences, four peer-reviewed journal papers and one international forum.

Contents

List of Abbreviations

2AFC	Two Alternative Forced Choice
AC	AutoCorrelation function
AR	AutoRegressive models
CAM	Color Appearance Model
CCD	Coefficient of Corralation of Distances
CD	Color Difference
CIE	Commission Internationale de l'Eclairage
CM	Co-occurrence Matrix
CRI	Carpet and Rug Institute
DMOS	Difference Mean Opinion Score
DWT	Discrete Wavelet transform
Eig	Eigenfilter
fps	frames per second
FFT	Fast Fourier Transform
Gb	Gabor filters
GLCM	Gray Level Co-occurrence Matrix
GM	Granulometry Moments
GMRF	Gaussian Markov random field
HVS	Human Visual System
ISO	International Organization for Standardization
ITU	International Telecommunication Union
LBP	Local Binary Patterns
LP	Laplacian pyramid
MAE	Mean Absolute Error
MME	Michelson's contrast Measure of Enhancement
MOS	Mean Opinion Score
PCC	Pearson Coefficient of Correlation
PSNR	Peak Signal to Noise Ratio
PVQ	user-Perceived Video Quality
PWC	Peli's Wavelet Contrast measure
PWD	Pseudo-Wigner distribution
RMSC	Root Mean-Squared Contrast
RMSE	Root Mean-Squared Error
SA	Spatial Activity
SE	Structuring Element
SME	Simple contrast Measure of Enhancement
SOVQM	Standard Objective Video Quality Metric

SP	Steerable Pyramid
SSIM	Structural Similarity Index Measure
SROCC	Spearman Rank Order Coefficient of Correlation
TA	Temporal Activity
TEM	Texture Energy Measures
VQEG	Video Quality Expert Group
VQM	Video Quality Metric
WME	Weber's contrast Measure of Enhancement

1

Introduction

The number of applications that rely on digital imaging as means of representing information continues to increase over the years. Since images and videos are typically intended to be viewed by humans, a considerable attention has been paid to image fidelity assessment: the objective assessment of the perceived differences between a reference (source) image and one or more corresponding test image samples.

Historically, the terms fidelity and quality have been used interchangeable in the image and video processing field, but they are often not the same. On the one hand, image fidelity assessment refers to quantifying perceptual differences between two samples (a reference and test sample). On the other hand, image quality assessment refers to assessing the subjective preference for one image over another. For example, Figure 1.1 shows an example were image fidelity assessment disagrees with image quality assessment. In Figure 1.1(a) the human subjects are asked to select from the two images the image that they prefer. In this case, the preference for the left side image is 54% from a pool of 15 human subjects, i.e., there is a mix-feeling between the observers of which image is the "better" image. In Figure 1.1(b) the human subjects are asked to select from the two images in the bottom, the image that is more similar to the top one. In this case; the image similarity between the top image and the left side image is 100% from the same pool of 15 human subjects. That is, the 15 human subjects selected the left side image as the more similar to the reference. Therefore, human observers may detect the differences between a reference image and its distorted (test image) version but this may not provide information about the human preference or quality. Additionally, fidelity and quality assessment are only equivalent when a reference or distortion free image is available in the assessment, e.g., the quality estimation of compressed video sequences is a task that is both fidelity and quality assessment because there is a reference video assumed to be distortion free or "perfect".

This dissertation researches the problem of measuring digital image fidelity, i.e., visual differences in images that an average human subject (observer) will perceive and report. The following four application areas and corresponding fidelity principles have been investigated:

(a) (b)

Figure 1.1: Comparison between image fidelity and image quality. (a) Image quality assessment vs (b) Image fidelity assessment.

- in the streaming of video sequences often is necessary to tune (de)coder and/or transmission parameters for their content for a target perceived video quality at the end-user device. In this area, this thesis proposes a real-time perceptual video quality measure for compressed sequences;

- in medical imaging (e.g. cardiac interventional x-ray), an image useful for the interventionalists is the one that has the perceived contrast ratio between the foreground and the background comparable (very similar) to that of the reference image. The reference image is an image where, according to the interventionalists, the diagnostically relevant details are presented under "ideal" detectability conditions. This thesis studies and develops contrast ratio measures for measuring perceived contrast changes in images;

- in the textile industry (e.g., assessment of appearance changes in textile floor coverings), texture of the textile materials is characterized to determine lifetime of textile products. This is typically done using digital imaging and texture analysis as a tool. This dissertation studies and evaluates texture analysis algorithms and the influence of their parameters

in the evaluation of appearance changes in texture;

- in multi-view imaging (e.g. live broadcasting), color correction is used to diminish color inconsistencies between views. There, the assessment of color differences is used to select the color correction algorithm that will produce the smallest perceived color difference between views. This dissertation proposes a novel color difference measure for natural scene color images.

1.1 Problem statement

As stated earlier, a considerable attention has been pay to the objective assessment of the perceived differences between a reference (source) and one or more test image samples, termed image fidelity assessment. In this area, the individual characteristics of the visual difference may vary from application to application. For instance, in image/video quality assessment of compressed sequences, the goal is to determine the global perceived differences between two images/videos. Typically one is a "perfect" image/video and the other has been subject to certain amount of distortion due to some process such as compression, transmission (Daly, 1992; Pappas et al., 2010). In cardiac interventional x-ray, image fidelity assessment often needs to quantify the visibility between a structure of interest such a vessel and its surrounding anatomical background, termed *contrast ratio*. In this case, the feature of interest is *image contrast ratio* and thus the image with the highest fidelity is the one with the smallest contrast ratio difference relative to a reference/standard sample (Kumcu et al., 2015a). In the textile industry, the evaluation of fidelity is based on the lifetime which is closely related to the surface appearance of the textile material. For example, appearance retention of textile materials is based on measuring changes in *color and texture* between the reference (new textile sample) and the test sample ("used" textile sample) (Aibara et al., 1999). *Color differences* (Fezza et al., 2014; Ly et al., 2015) are important in applications dealing with color quantization (Brun and Tremeau, 2002), color mapping (Morovic, 2008), among others.

Many of the image fidelity models in the state-of-the-art try to predict image fidelity using one-size-fits-all solutions based on sophisticated human visual system models which in general results in complex implementations. For instance, image fidelity measures often compute structural similarity between the reference and distorted images such as SSIM (Wang and Bovik, 2006) and its many alternatives improved from different perspectives (Gu et al., 2017, 2018; Ahar et al., 2018). Other approaches try to model the human visual system (HVS) using filter banks simulating attributes of perception such as contrast, graininess, sharpness and/or visual saliency (Zhang et al., 2014; Lissner et al., 2013; Larson and Chandler, 2010). Usually, the filter banks are fixed image filters inspired by human visual system models. However, more recent approaches have considered to learn the filter banks directly from the not distorted images

(reference images) (Guha et al., 2014).

More recently, many image fidelity measures rely on the computation of low level features (e.g., gradient statistics (Gao et al., 2018), statistics in the DCT domain (Li et al., 2017), among others) and their combination using machine learning algorithms (e.g., support vector machines (Yang et al., 2017), neural networks (Lukin et al., 2015), among others) to estimate the image fidelity scores (Gao et al., 2018; Ahar et al., 2018). Typically, the performance of these methods depends highly on the variety of image content included during the training phase. In other words, the main drawback of these methods is that an off-line training with enough samples representing the wide range of fidelity levels, extent of image details, color range and motion (for video sequences) is needed.

These fidelity measures are typically cumbersome for inclusion in any image processing algorithm or system (Wang and Bovik, 2006; Pappas et al., 2010). Also, it is typically more desirable to take advantage of the context or the individual characteristics of the visual differences depending on the application for achieving higher correlation with the differences perceived by the human observers. That is, to design application-specific fidelity assessment measures rather than a one-size-fits-all solution which in general are computationally complex to be included in any real time system and the gain in performance is usually not major compared with the application-specific models (Wang and Bovik, 2006; Pappas et al., 2010). In this thesis we focus on studying the application-specific fidelity assessment models and we demonstrate their advantages in applications intending to measure the visual differences that a human subject (observer) will perceive and report.

1.2 Contributions of this dissertation

We describe in the following paragraphs the contribution of this thesis to each of the aforementioned four image fidelity assessment tasks: (1) video quality estimation of compressed sequences, (2) evaluation of perceived contrast changes in digital images, (3) assessment of the appearance changes in texture, and (4) color difference estimation of natural scene color images.

Additionally, we have developed a software tool to compute fidelity assessment in images designed to assist image fidelity researchers providing easy access to a range of state-of-the-art measures which can be applied on a single pair of images and/or in a full database, as well as intuitive visualizations that aid data analysis, e.g., images and histograms of pixel-wise image differences, scatter plots and correlation analysis. This software tool is described in Appendix B.

1.2.1 Quality estimation of compressed video sequences

Quality estimation of compressed images/videos intends to measure the perceived image/video degradation of the compressed sample compared to the

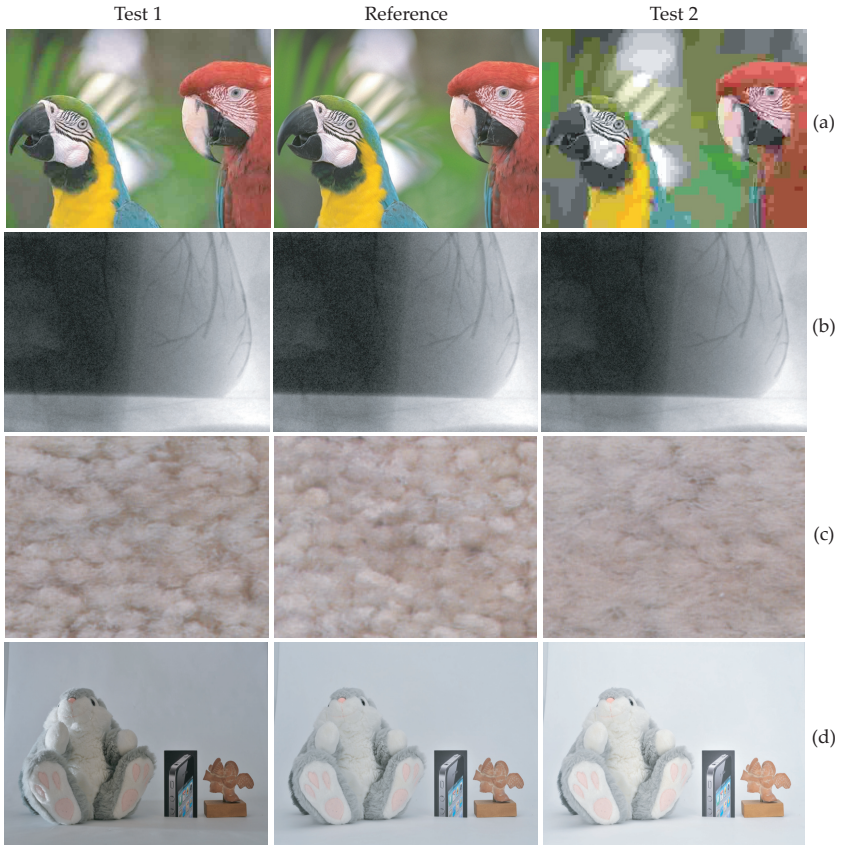

Figure 1.2: Image fidelity assessment examples: (a) quality evaluation of compressed images (Reference - uncompressed image, Test 1, 2 - images compressed with JPEG at two different bit rates); (b) contrast ratio changes estimation (Reference - image acquired at "optimal" radiation dose, Test 1, 2 - images acquired at two different lower radiation doses); (c) appearance changes assessment of global texture (Reference - unused textile floor covering, Test 1, 2 - textile floor coverings were subjected to physical degradation before digitizing); and (d) evaluation of color differences of natural scene color images (Reference - image acquired under "correct" illumination, Test 1, 2 - acquired under different illumination conditions).

ideal or perfect image/video. In general, the quality estimation of compressed video sequences is very useful for measuring distortions produced by compression and transmission errors where multiple artifacts are introduced like in the Figure 1.2(a). The test images are the result of compressing the reference image using different quantization parameter in JPEG compression. Image/Video quality rating is potentially applicable in digital TV or video streaming applications to tune (de)coder and/or transmission parameters for their content

for a target video quality at the end-user device. Image/Video quality rating can be used also for monitoring video quality at the end user device, e.g., for compliance reporting and quality control, among others (ITU, 1998; Video-Quality-Experts-Group, 2003).

This dissertation studies quality evaluation of compressed video sequences. We evaluate four of the most well-known state-of-the-art video quality measures on five different public video quality databases. Additionally, we propose a methodology to advance existing video quality measures by introducing video content related indexes in their computation. The performance of the proposed method is compared against their state-of-the-art counterparts. Our results show that unlike other conventional methods, the proposed method is of low complexity and satisfies the requirements of real-time applications (e.g., the proposed method runs at 75 fps for 720x380 pixels while the National Telecommunications and Information Administration General Model (Pinson and Wolf, 2004a) runs at 1 fps). At the same time, accuracy of the proposed method predictions is comparable with the conventional methods. This thesis studies quality evaluation of compressed video sequences in Chapter 3.

This research has been conducted within the framework of the "Telesurgery project - Digital operating room with live video feeds & real-time information at remote location" supported by "iMinds". The project involved the collaboration with academic as well as industrial partners including the research group for Media, Innovation and Communication Technologies (Ghent University), imec-MICT-UGent, the Department of Electronics and Informatics (Vrije Universiteit Brussel), imec-ETRO-VUB, and Barco. Part of this research has also been conducted within the framework of the "Panorama project - Ultra Wide Context Aware Imaging" of the ENIAC Joint Undertaking co-funded by grants from Belgium, Italy, France, the Netherlands, and the United Kingdom. The project involved the collaboration with industrial partners including Grass Valley Nederland B.V. and Bosch Security Systems.

We have proposed a video quality measure that is computational simple enough to be able to run in real time even without CPU/GPU optimization. Particularly, we have implemented a Python script able to compute perceived video quality at 12, 25 and 75 frames per second for 1920×1080, 1280×720 and 720×380 pixels, respectively. This software tool is described in Appendix C.

1.2.2 Evaluation of contrast ratio changes in images

Figure 1.2(b) shows an example of fidelity assessment where it is necessary to evaluate the contrast ratio changes. In the example from the Figure, image fidelity assessment needs to quantify the difference of the visibility of the coronary tree in the test image compared to the one in a standard reference image. Therefore, the differences are accounted by contrast ratio changes between the images. Fidelity assessment based on contrast ratio changes can be potentially used in image acquisition during interventional X-ray where specified areas of the acquired image are analyzed to determine the perceived contrast ratio difference between the test and a reference image. The reference image is an

image where the diagnostically relevant details (e.g., the coronary tree) are presented under "ideal" detectability conditions (Kumcu et al., 2015a). In multiview imaging, color correction (Fezza et al., 2014) and contrast enhancement techniques (Palma-Amestoy et al., 2009; Bertalmio et al., 2009) are often used to adjust the contrast, brightness and/or color settings of the cameras and/or displays. In these techniques, it is crucial to have an accurate measure of contrast ratio to produce images with the minimum perceived differences with respect to the reference, e.g., in a multi-camera system the contrast changes are measured with respect to the image acquired with the camera defined as the reference camera (Zhao et al., 2013).

This thesis investigates the evaluation of contrast ratio changes in images. We perform an extensive experimental evaluation based on a total of 6 image contrast ratio measures, each evaluated and tested on two image quality assessment databases. Also, we propose a novel methodology to compute contrast ratio in images by using local content analysis. The performance of the proposed method is compared against their state-of-the-art counterparts. The results show that the proposed method is able to accurately predict contrast decrements and increments better than the other state-of-the-art algorithms. Additionally, we test our methodology on a real case scenario (detection of changes in contrast level in interventional x-ray images acquired with varying dose). The results show that the proposed contrast ratio measure agrees with the subjective evaluation of interventionalists in interventional x-ray images. This thesis discusses the evaluation of contrast ratio changes in images in Chapter 4.

This research has been conducted within the framework of the "Panorama project - Ultra Wide Context Aware Imaging" of the ENIAC Joint Undertaking co-funded by grants from Belgium, Italy, France, the Netherlands, and the United Kingdom. The project involved the collaboration with academic as well as industrial partners including Philips Healthcare, the University of Leeds and the Ghent University Hospital.

1.2.3 Assessment of appearance changes in texture

Figure 1.2(c) shows an example of the evaluation of surface appearance changes in textile floor coverings. The most important visual parameter for this application is texture and therefore the images are compared in terms of global texture changes. Here, the texture pattern in the texture floor covering surface of the photograph of the image Test 1 is more similar to the reference than the texture pattern of the image Test 2. Some examples of fidelity assessment based on texture are wrinkling assessment (Na and Pourdeyhimi, 1995; Zhifeng et al., 2003), pilling assessment (Mendes et al., 2010) and assessment of appearance changes in textile floor coverings (Orjuela-Vargas, 2012), among others.

This thesis reviews and evaluates fourteen texture analysis descriptors for automated digital image-based evaluation of appearance changes in texture and discusses the impact of the parameter selection of the evaluated texture analysis techniques. We have studied the four texture descriptor categories: statistical

features, structural features, signal processing based features and model based features (Tuceryan and Jain, 1998; Zhang and Tan, 2002; Xie, 2008). The experimental evaluation is based on a total of three set of image databases. Our results show that the signal processing methods are the best performing with a strong correlation between the texture descriptors and human assessment in texture surfaces without complex patterns. The evaluation of global texture changes is studied in Chapter 5.

This research has been conducted within the framework of the "WEARTEX project - comparison of texture analysis techniques to assess WEAR labeling of TEXtile floor coverings" supported by "Ghent University" (WBS-element B/00565/01). The project involved the collaboration with academic as well as industrial partners including the vakgroep Textielkund (Ghent University) and the textile floor covering company LANO.

1.2.4 Evaluation of color differences in natural scene images

In color difference assessment of digital images, images are compared against a reference to determine if there are color inconsistencies between a test image and the reference. Figure 1.2(d) shows and example where the image reference was acquired under a "correct" light exposure and the two test images were acquired by using under and over exposure (image Test 1 and Test 2, respectively). The evaluation of color differences in images is used in color correction (Fezza et al., 2014; Ly et al., 2015), color quantization (Brun and Tremeau, 2002), color mapping (Morovic, 2008), among others.

This dissertation studies the evaluation of perceived color differences in natural scene color images. We review and evaluate eighteen state-of-the-art color difference measures as well as discusses their performances. The measures are tested in a total 25 different source images and three different color-related distortions. Additionally, we propose a novel method to compute color differences in natural scene color images based on the findings of the review. We base our measure on the fact that humans assess color difference in natural scene color images by comparing small patches. These patches are typically characterized for being homogeneous or for possessing an unique texture pattern. Our results show that the proposed color difference measure is able to predict changes of color more accurately than the other state-of-the-art algorithms. This thesis further discusses color difference assessment in Chapter 6.

This research has been conducted within the framework of the "Panorama project - Ultra Wide Context Aware Imaging" of the ENIAC Joint Undertaking co-funded by grants from Belgium, Italy, France, the Netherlands, and the United Kingdom. The project involved the collaboration with industrial partners including Grass Valley Nederland B.V. and Bosch Security Systems.

1.3 List of publications

The work developed in this thesis has been presented and discussed in six international conferences, four peer-reviewed journal papers and one international forum.

- The research in quality estimation of compressed video sequences has been published in one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2016b) and two conference proceedings (Ortiz-Jaramillo et al., 2014a, 2015c):

  – B. Ortiz-Jaramillo, J.O. Nino-Castaneda, L. Platisa and W. Philips, "Content-aware objective video quality assessment," Journal of Electronic Imaging, vol. 25, pp. 013011 1 - 16, 2016.

  – B. Ortiz-Jaramillo, A. Kumcu, L. Platisa and W. Philips, "Content-aware video quality assessment: predicting human perception of quality using peak signal to noise ratio and spatial/temporal activity," Proc. SPIE 9399, Image Processing: Algorithms and Systems XIII, pp. 939917 1 - 12, 2015.

  – B. Ortiz-Jaramillo, A. Kumcu, L. Platisa and W. Philips, "A Full Reference Video Quality Measure based on Motion Differences and Saliency Maps Evaluation," Proc. VISAPP, PANORAMA special session, vol. 2, pp. 714 - 722, 2014.

  Additionally, this research has been demonstrated at the Imec Technology Forum 2017 (cf. Appendix C) and it led to the ongoing collaboration discussion with two renowned industry players such as Telenet and Samsung.

- The studies performed on the evaluation of contrast ratio changes in images have been discussed in one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2018a) and two conference proceedings (Ortiz-Jaramillo et al., 2015b,a):

  – B. Ortiz-Jaramillo, A. Kumcu, L. Platisa and W. Philips, "Content-aware contrast ratio measure for images," Journal of Signal Processing: Image Communication, vol. 62, pp. 51 - 63, 2018.

  – B. Ortiz-Jaramillo, A. Kumcu, L. Platisa and W. Philips, "Computing contrast ratio in images using local content information," Proc. of the Symposium on Signal Processing, Images and Computer Vision, pp. 1 - 6, 2015.

  – B. Ortiz-Jaramillo, A. Kumcu, L. Platisa and W. Philips, "Computing contrast ratio in medical images using local content information," Proc. of the Medical Image Perception Conference, pp. 34 - 34, 2015 (abstract).

- The research realized in the area of assessment of appearance changes in texture has been presented in one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2014b) and one conference proceedings (Ortiz-Jaramillo et al., 2017)
  - B. Ortiz-Jaramillo, S.A. Orjuela-Vargas, L. Van-Langenhove, C.G. Castellanos-Dominguez and W. Philips, "Reviewing, selecting and evaluating features in distinguishing fine changes of global texture," Pattern Analysis and Applications, vol. 17, pp. 1 - 15, 2014.
  - B. Ortiz-Jaramillo, L. Platisa and W. Philips, "iFAS: Image Fidelity Assessment," Proc. International Workshop on Computational Color Imaging, pp. 83 - 94, 2017.
- The work proposed in the evaluation of color differences in natural scene color images has been reported in one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2018b) and one conference proceedings (Ortiz-Jaramillo et al., 2016a):
  - B. Ortiz-Jaramillo, A. Kumcu, L. Platisa and W. Philips, "Evaluation of color differences in natural scene color images," Journal of Signal Processing: Image Communication, Submitted 2018.
  - B. Ortiz-Jaramillo, A. Kumcu and W. Philips, "Evaluating color difference measures in images," Proc. of international Conference on Quality of Multimedia Experience, pp. 1 - 6, 2016.

1.4 Organization of this dissertation

The remaining of this dissertation is organized as follows. In Chapter 2 we introduce the background theory about image fidelity assessment including subjective and objective evaluation. Also, we revise the methodology to evaluate numerical fidelity assessment measures. Additionally, we propose a software tool designed for easy experimentation with the studied state-of-the-art image fidelity methods, including elaborate data analysis and evaluation of image fidelity measures.

In Chapter 3 we study the quality evaluation of compressed video sequences. We study and evaluate four of the most well-known state-of-the-art video quality algorithms on five different public video quality databases. Additionally, we propose a method to advance existing numerical video quality measures by introducing *content related indexes* in their computation.

This thesis reviews in Chapter 4 the methods for computing the contrast ratio in images. In the same Chapter, we propose a measure to compute contrast ratio in local image patches. Also, we perform an extensive experimental evaluation based on a total of six image contrast ratio measures, each tested on two image databases and we test our methodology on predicting subjective evaluation of interventionalists in interventional X-ray fidelity assessment.

We survey in Chapter 5 the evaluation of texture in images. We review and evaluate features in the evaluation of appearance changes in texture. We evaluate fourteen texture descriptors for characterizing changes in texture due to degradation. We include descriptors based on statistics, filtering, structural and models. Also, we discuss the impact of the parameter selection of the evaluated texture analysis techniques.

We investigate the color related aspect of image fidelity assessment in Chapter 6. We study eighteen state-of-the-art color difference measures and we discuss their performances. The measures are tested on one public available database and three different types of color-related distortions. Additionally, we propose a novel method to compute color differences in natural scene color images based on the findings of the review.

This thesis concludes in Chapter 7 where conclusions over the related areas are drawn and possible directions for future work are discussed.

2

Image fidelity assessment

2.1 Introduction

Historically the terms fidelity and quality have been used interchangeable in the image and video processing field, but they are often not the same (Silverstein and Farrell, 1996; Wang and Bovik, 2006; Pappas et al., 2010). On the one hand, image/video fidelity assessment refers to the ability to quantify visual differences between two samples (a reference and test sample), in other words how close an image/video is to a given reference (Pappas et al., 2010). On the other hand, image/video quality assessment refers to the preference for one image/video over another (Silverstein and Farrell, 1996).

For example, Figure 2.1 shows the comparison of two color images to a reference sample. This comparison could be performed by asking two different questions to a human observer: in the case of quality assessment, the observer would need to answer the question which image do you prefer the right side or the left side image? (Figure 2.1(a)) while in the fidelity assessment case the observer would need to answer the question which image is more similar to the reference one (top image)? (Figure 2.1(b)). Note that the similarity score is very high for the left side image (80% of the observers find this image more similar). However, the human observers overall prefer the right side image (85% of the observers prefer this image) over the left side image (overall preference equal to 15%). Therefore, human observers may detect the differences (fidelity) between a reference image and its processed (test image) version but this may not provide information about the human preference (quality) between the images. In general, although fidelity and quality are closely linked for many tasks these two measures highly disagree on the context and/or task at hand (Silverstein and Farrell, 1996; Pappas et al., 2010).

In this thesis, we study computer algorithms to quantify the visual differences typically perceived and reported by human observers. We deal with application-tailored fidelity assessment tasks such as, quality estimation of compressed video sequences (Chapter 3), contrast difference evaluation of digital images (Chapter 4), assessment of appearance changes in texture (Chapter 5) and color difference assessment of natural scene color images (Chapter 6). In

Figure 2.1: Example of a paired comparison setup for a natural scene color image. (a) Image preference score based on 15 observers for the left side image is 15% and for the right side image is 85%. (b) Image similarity score based on 15 observers for the left side image is 80% and for the right side image is 20%.

general, we differentiate between the terms fidelity and quality assessment, except for Chapter 3 where the quality estimation of compressed video sequences is designed to assess the perceived/perceptual difference between the test video and its reference ("perfect"). That same problem is modeled by many existing algorithms typically referred to as "full reference perceived video quality measures", which corresponds to the paradigm of image fidelity assessment; accordingly, in order to comply with the common terminology, we will refer to this particular model as quality assessment. Since the goal is to predict what a human observer would perceive and report, the simplest solution is to subjectively evaluate using human observers according to certain well-defined criteria (See Section 2.2 for details). However, subjective evaluation is in general complex, expensive, time consuming and therefore unpractical for real time image/video processing. Thus, many researchers have proposed numerical methods for predicting fidelity from the image/video data. In any case, subjective evaluation is currently considered the benchmark for any given image fidelity assessment task.

The purpose of this Chapter is to introduce the background theory concerning image/video fidelity assessment and to define the methodology to evaluate objective fidelity assessment measures. The rest of this Chapter is organized as follows. In Section 2.2 we describe subjective fidelity assessment. Section 2.3 studies the performance evaluation of numerical fidelity assessment measures. Later, we propose a software tool designed to assist numerical image fidelity evaluation in Section 2.4. Finally, we summarize in Section 2.5.

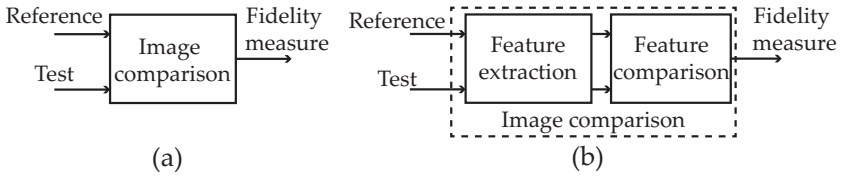

Figure 2.2: General framework for image/video fidelity assessment. (a) observer-based fidelity assessment, (b) feature-based image fidelity assessment.

2.2 Subjective fidelity assessment

Since the end goal is to measure the visual differences typically perceived and reported by human observers, subjective evaluation represents the benchmark for image/video fidelity assessment (ISO-10361:2000, 2005; ITU, 1998). In subjective fidelity assessment, image/video samples are compared to their respective reference (source images/videos) by human observers. The result of such an evaluation is a subjective score per evaluated test image/video sample. The subjective assessment is the most well-known and most widely used technique for measuring fidelity. Fidelity assessment can be done by assessing observer-based fidelity or feature-based fidelity where the features of interest are typically determined by the use case/application. Therefore, different applications require different testing procedures.

Figure 2.2(a) shows the framework for observer-based fidelity assessment where the observers are asked to assess how similar the test sample is to the reference without indicating the individual characteristics of the visual differences. Additionally, the test images could exhibit multiple perceived distortions/artifacts. Figure 2.2(b) shows the framework for feature-based image fidelity assessment. In this case, the observers know the nature of the differences. That is, the observer has to evaluate the visual differences in terms of certain well defined feature, e.g. color, texture, contrast.

Subjective scores are typically collected by means of psycho-physics because psycho-physics provides the tools for quantifying visual differences/similarities. In psycho-physics there are two main types of tasks to collect subjective scores, namely adjustment and judgment (Winkler, 2005; Kingdom and Prins, 2010a). The former includes tasks such as setting the threshold amplitude of a stimulus, canceling an image difference, or matching a stimulus to a given reference. The latter includes forced choices between alternatives and magnitude estimation on a rating scale (Winkler, 2005). We do not intend to fully study psycho-physics but instead describe the methodologies that are commonly used in the subjective fidelity assessment field. Note that the adjustment methods are more easily treated in a signal detection framework than for an image similarity framework. Since we are interested in measuring visual differences (similarities) typically perceived and reported by human observers, all the methods discussed in this thesis are judgment based methodologies.

In general, the standard way of measuring perceived differences is based on subjective testing performed by appropriately chosen human observers. The human observers can be expert or "naive" observers depending of the task at hand. For instance, the contrast difference assessment between x-ray images requires expert observers evaluation while the quality evaluation of compressed video sequences can be performed by "naive" observers. Afterwards, a set of well selected images/videos samples is shown to the human observers following well-defined rules, for example, the ITU-R BT.500-11 (ITU, 1998) for multimedia applications and ISO 10361 standard for textile floor coverings (ISO-10361:2000, 2005).

Although subjective assessment is the most well-known and most widely used technique for assessing fidelity of image/video-based systems (ITU, 1998; ISO-10361:2000, 2005), such a technique is in general complex, expensive, and time consuming. Therefore, it is unpractical for real time image/video processing and hard to incorporate into a system design process. For this reason, many researchers have proposed objective (numerical) methods for predicting fidelity directly from the image/video data. However, annotated image/video databases with subjective fidelity ratings are essential ground truth for developing, training, testing, and benchmarking algorithms for objective fidelity assessment. Thus, we describe the most well know and widely used procedures to collect subjective scores in the rest of this Chapter. Objective fidelity assessment and related contributions of this thesis are discussed in detail for quality estimation of compressed video sequences, evaluation of contrast changes in digital images, assessment of appearance changes in texture and color difference estimation of natural scene color images in Chapters 3, 4, 5 and 6, respectively.

2.2.1 Subjective assessment of images

In still image fidelity assessment, two or three images are displayed simultaneously to be assessed (compared) by human observers. The observers are typically asked to assess the similarity of the test images to the reference. Then, the subject score is expressed with a grading scale system (Ponomarenko et al., 2015). For instance, observers might be asked to rate the amount of difference between two images, like in Figure 2.3(a), using a given scale, e.g., a five levels rating scale: "Imperceptible", "Small", "Medium", "Obvious" and "Very Obvious" difference. When presenting three images (the reference and two test images) to the observer, e.g., see Figure 2.3(b), the task of the observer is to select the image with the lowest perceived difference compared to the reference (Ponomarenko et al., 2015).

In image fidelity assessment, subjective evaluation is typically performed as follows:

- A set of source images is collected with the purpose of covering various image content (extent of texture, color content). Figure 2.4 shows 4 images from the Kodak set which is a typical example of distortion-free (source) images (KODAK, 2013).

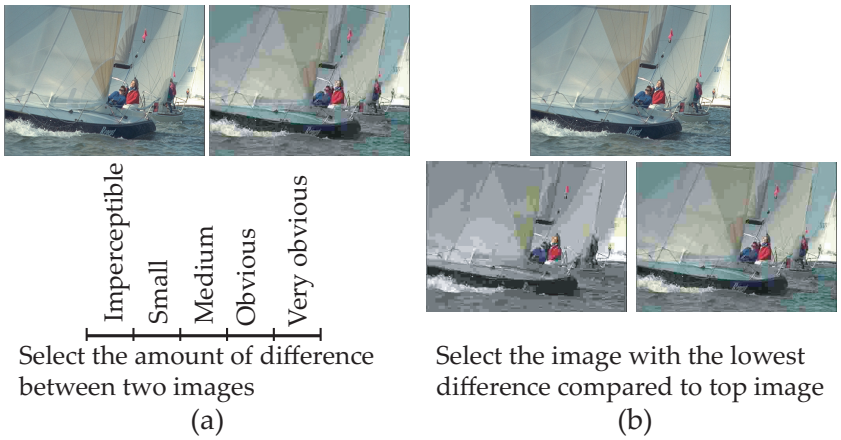

Figure 2.3: Presentation of the test images based on the rating method, (a) judgment is expressed with a grading scale, (b) judgment is based in pair comparison.

Figure 2.4: Images from the Kodak set.

- Each source image is processed/manipulated depending on the feature of interest. For instance, the images could be compressed using JPEG, color mapped, transmitted over a noisy channel, among others; all of them producing visual differences. Figures 2.5(a), (b) and (c) show examples of processed images to be used in a fidelity assessment database (JPEG compression, contrast change and mean shift, respectively) (TID2013, 2013).

- Human observers evaluate the visual differences between the source images and their respective processed images (Figure 2.6). Each observer assigns a fidelity score to the given test image or select the image with the lowest difference. Note that this set of techniques also apply for medical applications. However, for medical applications the human observers should be medical experts (e.g., x-ray interventionalists).

- One score per sample is computed from human observers representing the average fidelity of the test image. The final score is computed by averaging the scores of all the scores collected or by summing-up the number of times each test image was selected as the most similar to the

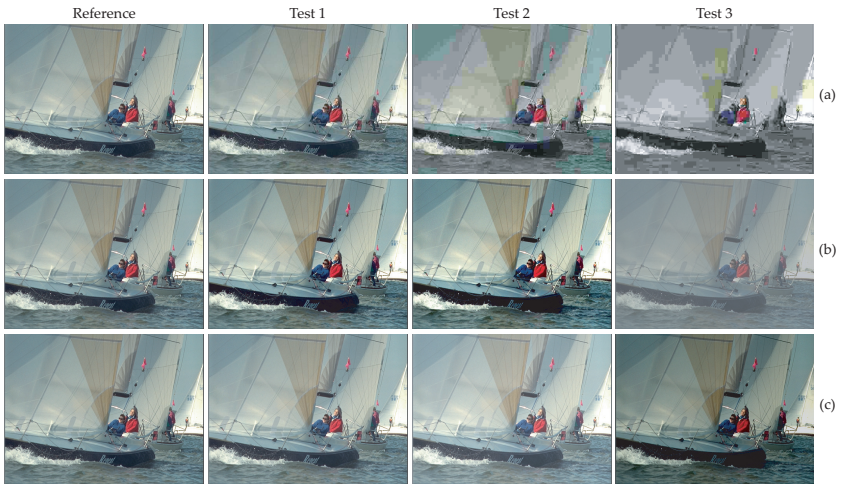

Figure 2.5: Images processed using (a) JPEG compression, (b) contrast change and (c) mean shift.

Figure 2.6: Human observers for multimedia applications.

reference. Often the average is accompanied by the standard deviation of the observers' scores. The result is a set of images with an assigned fidelity subjective score (see Figure 2.7).

This type of databases, termed image fidelity/quality databases, are intended for the evaluation of numerical image fidelity assessment measures. The databases tested in this thesis are described in Appendix A.

2.2.2 Subjective assessment of videos

The subjective fidelity (quality) assessment of image sequences (videos) has been described in the ITU recommendations (Winkler, 2005). For video sequences two types of approaches are adopted for the displaying of the test sequences. The first one uses parallel displaying where the human observers are watching two sequences at the same time: one is the reference, the other

Figure 2.7: Example of mean opinion scores assigned to JPEG distorted images. Mean opinion score (ranging 0-1) (a) 0.72, (b) 0.63 and (c) 0.30.

one is the distorted sequence (ITU, 2012). The other methodology involves displaying each video sequence independently. That is, the source sequence is first shown to the observer followed by the distorted version and/or vice versa (Li et al., 2014; Van-Wallendael et al., 2016). Some methodologies use what it is defined by ITU as hidden reference where the observer judges each sequence individually and the reference sequence is often mixed within the processed video sequences.

- Analogous to image fidelity assessment, a set of source video sequences (distortion free or "perfect") is collected with the purpose of covering various video content characteristics (extent of texture and motion, color content).
- Afterwards each source video sequence is processed for the purpose of creating the test video samples. In video quality databases, the "perfect" video sequences are usually compressed, e.g., using H.264, MPEG2, or corrupted by introducing artificial transmission or network errors.
- Human observers evaluate the visual differences. This evaluation can be performed by using one of the following displaying methodologies: (a) double stimulus where the reference is independently presented before or after the distorted sequence and (b) single stimulus where only the distorted sequence is presented to the observer or the source sequence is first shown to the observer followed by the distorted version and/or vice versa. The observers separately rate the two sequences (if necessary) on a continuous or discrete quality scale (Figure 2.8 shows these scales).
- The average from all the observers is collected as the overall fidelity (quality) of the processed sequence. Often the average is accompanied by the standard deviation of the observers' scores.

This type of databases, termed video quality databases, are intended for the evaluation of numerical video quality measures.

2.2.3 Subjective assessment of surface appearance

In the textile industry, the evaluation of appearance consists in visually identifying deviations from a reference sample. Standards for evaluating the ap-

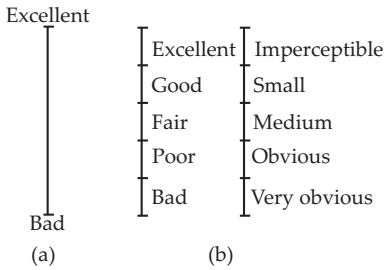

Figure 2.8: Rating scale. (a) Continuous rating from "bad" to "excellent" (typically scaled in a range 0-5 or 0-100), (b) discrete rating scales for absolute category rating and degradation category rating (left and right, respectively).

Figure 2.9: The most common devices for introducing degradation in textile floor coverings. (a) castor chair, (b) Vetterman and (c) Hexpod.

pearance retention of textiles are designed using a set of samples exhibiting transitional degrees of degradation due to daily use. The evaluation is typically performed by a panel of certified experts (ISO-10361:2000, 2005). In general, the evaluation is independent of the textile material type and performed as follows:

- A set of textile samples is built simulating daily exposure by using mechanical devices, e.g., the castor chair, Vetterman and Hexpod are the most well-known devices for introducing degradation to the textile surface of textile floor coverings (ISO 4918:2016, 2016; ISO-10361:2000, 2005) (Figures 2.9(a), (b) and (c), respectively).
- A panel of certified experts evaluate the different visual characteristics of the textile (see Figure 2.10). A score ranging from 1 to 5 is assigned to the textiles, where a severe change is evaluated with the score 1 and no perceived difference is evaluated with the score 5 (ISO-10361:2000, 2005).
- The average rate (or score) is collected from the panel of observers representing the overall appearance of the textile material. The result is a set of textile samples with an assigned appearance change, termed, reference scales (see Figure 2.11).

Figure 2.10: Panel of certified experts evaluating different textile materials.

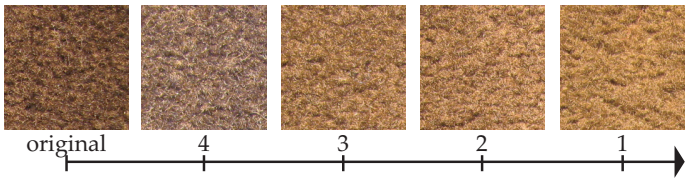

Figure 2.11: Changes due to degradation in a textile floor covering. Here the full set of wear levels in steps of 1.

This type of databases are typically certified by international standards and composed of physical samples used to assess new textile floor coverings and/or to train human inspectors (ISO 9405:2015, 2015). Currently, the evaluation of new degraded textile samples is done by comparing the texture characteristics of the degraded sample to the characteristics of the certified physical reference scales. However, a major drawback is that the physical characteristics of the certified reference scales can change over time while being exposed to involuntary damage (Orjuela-Vargas, 2012). Therefore, an alternative has been proposed based on photographs of the certified physical reference scales. These digital reference scales are proposed by the Carpet and Rug Institute (CRI) (CRI Test Method 103, 2015) while the physical reference scales are proposed by the International Organization for Standardization (ISO) (ISO 9405:2015, 2015). Additionally, the use of photographs of the physical reference scales allows the use of computational technologies for the objective assessment of appearance changes in textiles and therefore they can be used as well for the evaluation of numerical fidelity measures. We devote Chapter 5 to study this topic from the image processing point of view.

In practice, the assigned appearance change is used for validating the textiles under inspection. For instance, for commercial application in the Vetterman and Hexapod tests, the visual inspectors give scores from 1 to 5 for samples exposed to 5000 and 22000 rotations in the drums. The final score is a combination of the two evaluations according to the formula

$$\text{score} = 0.75 \times \text{score}_{5000} + 0.25 \times \text{score}_{22000},$$

where score_x is the appearance change assigned to the textile floor covering after being exposed to x rotations using the Vetterman or Hexapod tests. A score of

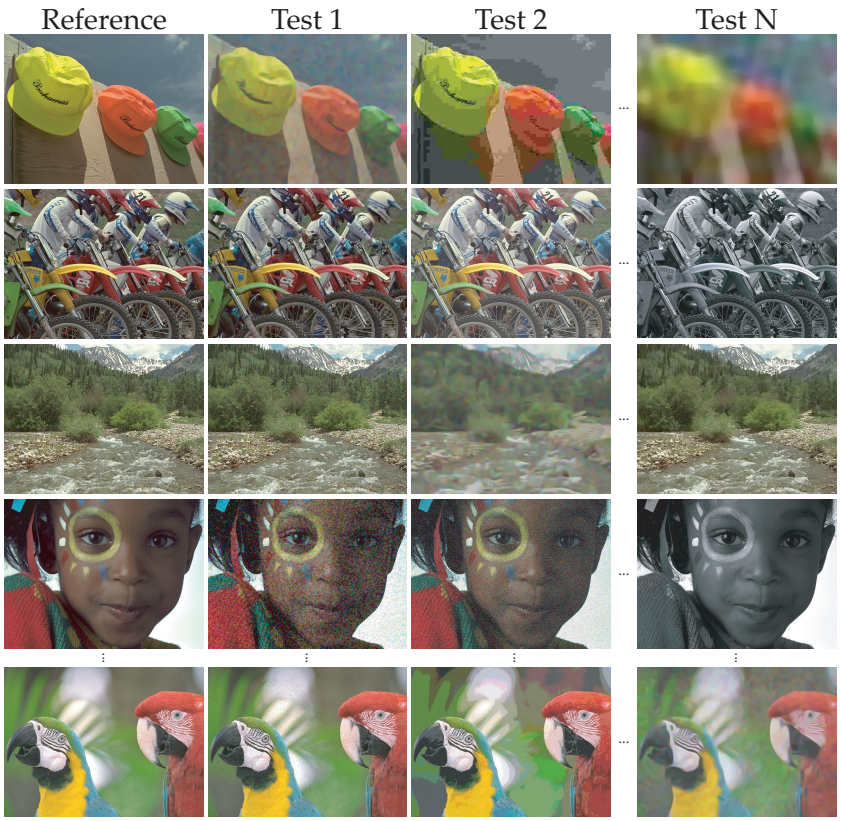

Figure 2.12: Examples of source and distorted images from TID2013 database.

2 or more is a pass and a result of 2.4 or more is a pass for intensive use (ISO-10361:2000, 2005). The current reference samples used by the standards and some of the construction details are listed in Appendix A.4.

2.3 Evaluation of objective fidelity assessment measures

Section 2.2 showed that image/video fidelity databases are intended for evaluating the performance of numerical fidelity measures. These databases are normally composed of a number of source images (references) and a number of test (distorted or processed) images/videos. Each source image/video has its corresponding test images/videos.

Figure 2.12 shows an example of typical images from an image quality database (the reader can find details about this database in Section A.1). A

subjective image/video quality databases is a set of $M \times (N+1)$ images where M is the number of source images and N is the number of distorted/test images per source sample. Additionally, each distorted image has associated a subjective score assigned by subjective evaluation. This set of data is intended for the evaluation of numerical video/image fidelity measures. In general, the numerical fidelity measures use the image data to estimate the visual differences perceived and reported by the human observers.

2.3.1 Benchmarking of numerical fidelity assessment measures

The performance of fidelity assessment measures is estimated by comparing the predicted numerical fidelity values with the reported human scores, i.e., various indices are evaluated between subjective scores and the numerical estimations. A common and well accepted way of benchmarking numerical fidelity assessment measures is to evaluate the following indices:

- Pearson Coefficient of Correlation (PCC) (Chen and Popovich, 2002a) is used to measure the accuracy of the tested fidelity measure using a linear model. The PCC between two discrete feature vectors $\mathbf{x}$ and $\mathbf{y}$ of n elements is defined as

$$\mathrm{PCC} = \frac{\sum_i^n (x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}})}{\sqrt{\frac{\sum_i^n (x_i - \mu_{\mathbf{x}})^2}{n}} \sqrt{\frac{\sum_i^n (y_i - \mu_{\mathbf{y}})^2}{n}}},$$

  where, x_i and y_i are the ith element of $\mathbf{x}$ and $\mathbf{y}$, respectively. $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ are the average of $\mathbf{x}$ and $\mathbf{y}$.

- Spearman Rank Order Coefficient of Correlation (SROCC) (Chen and Popovich, 2002b) is a measure of the monotonicity of the tested fidelity measure. The SROCC is defined as

$$\mathrm{SROCC} = 1 - \frac{6 \sum_i^n (\mathrm{R}x_i - \mathrm{R}y_i)^2}{n^3 - n},$$

  where $\mathrm{R}x_i$ and $\mathrm{R}y_i$ are the ranks of the ith element of $\mathbf{x}$ and $\mathbf{y}$, respectively.

- Coefficient of Correlation of Distances (CCD) (Székely et al., 2007) measures the accuracy for non-linear models. The CCD is defined as

$$\mathrm{CCD} = \frac{1}{n^2} \sum_i^n \sum_j^n A_{ij} B_{ij},$$

  where $A_{ij} = a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}$ and $B_{ij} = b_{ij} - b_{i\cdot} - b_{\cdot j} + b_{\cdot\cdot}$ are the distance matrices of the data. Each element of the distance matrices is defined as $a_{ij} = \|x_i - x_j\|$ and $b_{ij} = \|y_i - y_j\|$. Here, $a_{i\cdot}$ and $b_{i\cdot}$ are the

ith row average of the distance matrices; $a_{\cdot j}$ and $b_{\cdot j}$ are the jth column average of the distance matrices; and $a_{\cdot\cdot}$ and $b_{\cdot\cdot}$ are the grand average of the distance matrices.

In these measures, accuracy refers to the ability to predict the subjective fidelity scores with low error. This aspect is measured by using the PCC for linear models and CCD for non-linear models. The monotonicity is the degree to which predictions of the model agree with the magnitudes of subjective quality scores (Video-Quality-Experts-Group, 2003). This aspect is measured with the SROCC.

We use the rule of the thumb for interpreting the size of a correlation coefficient (Mukaka, 2012), i.e., we use the following descriptive scale:

Size of Correlation	**Interpretation**
0.90 to 1.00	Very strong correlation
0.70 to 0.90	Strong correlation
0.50 to 0.70	Moderate correlation
0.30 to 0.50	Weak correlation
0.00 to 0.30	Very weak correlation

We use only correlations to compare the rankings encoded in the scores given by the human observers and the numerical fidelity measures. Additionally, we use the Fisher's z transform defined as

$$z' = 0.5 \log \left(\frac{1 + \text{correlation}}{1 - \text{correlation}} \right),$$

with the purpose of comparing correlation coefficients on a linear scale (Borenstein et al., 2009). The percentage increase of a method A compared to a method B using the z' values is computed as

$$100 \frac{z'_{\mathrm{A}} - z'_{\mathrm{B}}}{z'_{\mathrm{B}}}.$$

Some applications require predicting the expected error (see Chapter 3), i.e., the expected difference between the subjective scores and the numerical estimations. To do so the root mean-squared error (RMSE) and the Mean Absolute Error (MAE) are used. The RMSE and MAE are defined as follows

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i^n (x_i - y_i)^2}{n}}$$

and

$$\mathrm{MAE} = \frac{\sum_i^n |x_i - y_i|}{n},$$

respectively. While RMSE is the degree to which the model maintains prediction accuracy over a range of different test samples, the MAE is a measure of the expected error of a new sample, i.e., the expected difference between the subjective and objective assessment.

Previous indexes have two major disadvantages: (1) they do not consider the variability of the reported human scores and (2) these indexes assign uniform weights to all fidelity levels albeit in many applications high-fidelity images are usually more important than the images displaying big fidelity differences (Sheikh et al., 2006; Winkler, 2009; Hobfeld et al., 2011; Wu et al., 2018). To solve these issues more advanced statistical techniques have been proposed. In this manuscript we outline the main principles of these techniques as we believe they will bring valuable insights in the continuation of our work but we do not include them in our experiments.

In (Sheikh et al., 2006), in addition to the classical SROCC and PCC indexes, two statistical hypothesis tests are used for the benchmarking of image quality measures. The first one assumes Gaussianity of the residual differences between the quality predictions and subjective scores and uses the F-statistic for comparing the variance of two different image quality measures. The goal of the test is to determine whether the residuals of the two image quality measures are statistically indistinguishable, that is, if there are statistical differences between the two measures.

The second hypothesis test used in (Sheikh et al., 2006) considers the variability of the reported human scores. This hypothesis test is based on the individual quality scores. This method proposed in (Video-Quality-Experts-Group, 2003) compares image quality measures against the individual subjective human scores. The residuals of the individual models are used to indicate if two image quality measures are statistically different. This method typically discriminates between image quality performances better than when comparing correlation coefficients because it considers the variability of the reported human scores by using the raw data.

More recently in (Wu et al., 2018), the authors study the problem of assigning uniform weights to all fidelity levels in the benchmarking of image quality algorithms. They explore a weighted rank correlation index which weights higher correctly ranked high-quality images and weights lower insensitive rank mistakes (two perceptually similar images). That is, this measure assigns its weights by both the quality level and the variability of the reported human scores.

2.3.2 Multiple statistical comparisons

Statistical tests for multiple comparisons are numerical methods for statistical inferences. The objective of this type of tests is to determine if there exist statistically significant differences in performance between the compared numerical fidelity measures. Here, the data (performance) is given in terms of correlations. Particularly, we use the test based on Friedman ranks (Garcia et al., 2010b). We explain this test by using an example. We compare the performance of four video quality assessment algorithms in terms of PCC in six video quality databases (see Chapter 3 and Appendix A.5 for details about the test data and the compared algorithms).

Table 2.1 shows the individual performance of the compared video quality

Table 2.1: Performance comparison of four different methods (PSNR, CPSNR, SOVQM and VQAD) for six public video quality databases. Performance is given in terms of PCC. The ranks in the parentheses are used in the computation of the test.

Database	PSNR	CPSNR	SOVQM	VQAD
IRCCyN	0.57 (4)	0.81 (2)	0.85 (1)	0.79 (3)
IVP	0.69 (4)	0.91 (1)	0.90 (2)	0.86 (3)
IRCCyN ic	0.83 (4)	0.89 (3)	0.94 (1)	0.92 (2)
CIF EPFL-PoliM	0.60 (4)	0.86 (3)	0.93 (1)	0.91 (2)
4CIF EPFL-PoliM	0.67 (4)	0.89 (2.5)	0.89 (2.5)	0.95 (1)
LIVE	0.56 (4)	0.81 (1.5)	0.73 (3)	0.81 (1.5)
Average rank (R)	0.65 (4)	0.85 (2.1)	0.87 (1.7)	0.87 (2.1)

measures. The test statistic for the multiple statistical comparisons based on Friedman test is computed as

$$z_{ij} = (R_i - R_j)\sqrt{\frac{k(k+1)}{6n}},$$

where $k = 4$ is the number of algorithms and $n = 6$ is the number of databases. R_i and R_j are the average rankings over all databases for algorithms i and j, respectively. Afterwards, the z value is used to find the corresponding probability (p-value) from the Normal distribution table (Garcia et al., 2010b). Thereafter, the p-values are adjusted using Bonferroni - Dunn correction with the purpose of taking into account that multiple test comparisons are conducted (Dinno, 2015). The Bonferroni - Dunn correction is defined as $\min(1, (k-1)p)$ where p is the p-value. The obtained p-value is a number between 0 and 1 and used to determine if an algorithm performs statistically better than another. The p-values are interpreted in the following way:

- a small p-value (typically ≤ 0.05) indicates strong evidence against the hypothesis that the compared algorithms perform equally well in terms of correlation;
- a large p-value (> 0.05) indicates weak evidence against such a hypothesis therefore the compared algorithms are likely to perform equally well in terms of correlation.

Table 2.2 shows the adjusted p-vales for the performance comparison of Table 2.1. Note that this example only includes six databases. However, many studies include multiple statistical comparisons with $n \gg 6$. In such a case, displaying the data as a table is difficult to read and interpret. Therefore, we use an intuitive visualization of the data. We use the box plot which is a graphical representation of the data through their quartiles (Massart et al., 2005). Box plots also have lines, termed whiskers, indicating the variability outside the upper and lower quartiles. Outliers are typically plotted as individual points. Note that the box plots and the multiple statistical comparisons

Table 2.2: Multiple statistical comparisons based on Friedman test. For the entries above the main diagonal, the direction of the arrow indicates which of the two video quality measure (row or column) performs better. The equal sign stands for a tie. Elements below the main diagonal are p-values. If the p-value is higher than 0.05, there were no significant differences between the video quality measures.

Model	PSNR	CPSNR	SOVQM	VQAD
PSNR	–	↑	↑	↑
CPSNR	0.041	–	=	=
SOVQM	0.007	1	–	=
VQAD	0.030	1	1	–

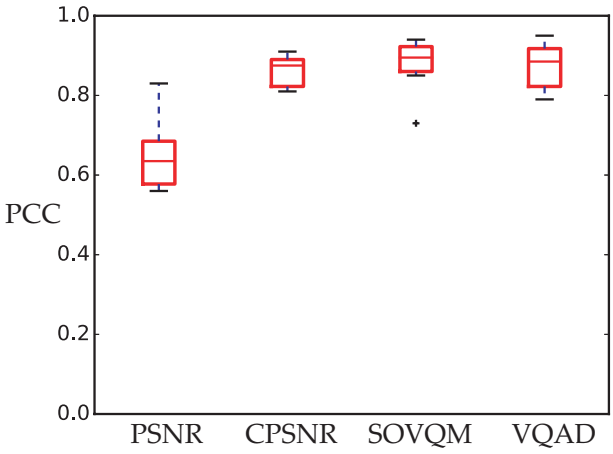

Figure 2.13: Performance comparison of the considered video quality measures. The box plot was created using the six correlation coefficients in Table 2.1.

based on Friedman test are non-parametric, i.e., they display and test variation in samples without making any assumptions of the statistical distribution of the data.

Figure 2.13 shows the box plot representation of the data in Table 2.1. That is, the box plot shows the variability in performance across different databases, i.e., how well the tested measures perform across different data type. This graphical representation is used together with the multiple statistical comparisons based on Friedman test to draw conclusions over the test data. In our example, from the multiple statistical comparisons and the box plots we found that there are no statistical significant differences in terms of performance between CPSNR, SOVQM and VQAD. Also, we found that CPSNR, SOVQM and VQAD outperform PSNR.

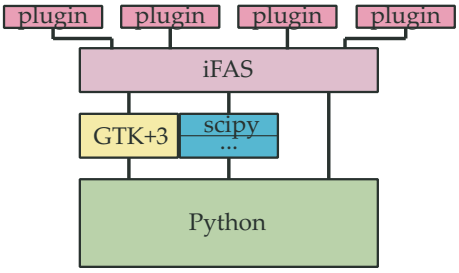

Figure 2.14: Simplified iFAS structure.

2.4 Image fidelity assessment software

Clearly there is a need for application-tailored fidelity assessment measures rather than a one-size-fits-all solution. While many candidate numerical fidelity measures already exist, testing them and identifying the best ones for a given use case is far from easy. Thus, it often takes a considerable amount of time and effort to even prepare the test environment for benchmarking, validation and/or developing. To the best of our knowledge, only few related software packages are available and they have limited features/functionality and/or are not freely available. For instance, Krasula et. al. proposed one Matlab based interface for testing 8 well-known image quality measures (Krasula et al., 2011). However, the interface is not easy to extend for new fidelity measures and it has no mechanism of benchmarking, correlation analysis or model analysis. Murthy and Karam developed IVQUEST which is the most complete open source user interface for subjective and objective image quality evaluation as well as correlation analysis (Murthy and Karam, 2010). Additionally, the interface allows easy extensions by writing pieces of Matlab code. However, the benchmarking and correlation analysis are limited to linear correlation analysis. Also, the interface is limited to 15 general image quality measures. Therefore, even though these platforms are very useful to test image quality measures, they are very limited in scope and available methods. Furthermore, they have been implemented in Matlab, which is a non-free platform.

In this thesis, we seek to alleviate such problems. We have developed a software tool "iFAS: (image Fidelity Assessment)" and made it freely available for non-commercial use. Figure 2.14 shows the simplified iFAS structure. iFAS structure is based on Python (Python, 2016) and depends on GTK+3 (GLib, 2016) as well as third party libraries for the implementation of the user interface, the mathematical models and plugins. Plugins are Python modules for extending the functionality of iFAS. The third party libraries are Scipy 0.17.0 together with its core packages, particularly, NumPy 1.11.0, Matplotlib 1.5.1, pandas 0.17.1, nose 1.3.7, Cython 0.25.1 and Scikits 0.12.3 (SciPy, 2016); PIL 1.1.7 (PIL, 2016); PyGObject (aka PyGI) 3.20.0 (GLib, 2016); Pycairo 1.10.0 (Pycairo, 2016); PyWavelets 0.5.0 (PyWavelets, 2016). iFAS pro-

vides the following image fidelity assessment tools:

- Single Source - Single Sample: this type of analysis computes a set of numerical fidelity measures between one reference image and one corresponding test image.
- Single Source - Multiple Sample: this type of analysis computes a set of numerical fidelity measures between one reference image and a number of corresponding test images.
- Multiple Source - Multiple Sample: this type of analysis computes a set of numerical fidelity measures between a number of reference images and a number of test images.

iFAS includes a set of 44 state-of-the-art fidelity measures including 18 color difference measures for digital images (see Chapter 6), 14 texture analysis algorithms (see Chapter 5), 6 image contrast ratio measures (see Chapter 4) and 6 image quality measures (peak signal to noise ratio, structural similarity index measure (Zhou et al., 2014), blocking (Muijs and Kirenko, 2005), noise estimation (Goossens, 2006), blurring (Platisa et al., 2011) and edge peak signal to noise ratio (Lee et al., 2009)).

iFAS allows visualization of scatter plots between two fidelity measures or a fidelity measure and corresponding subjective scores. Additionally, iFAS computes the following correlation indexes typically used as performance indicators of fidelity measures: PCC (Chen and Popovich, 2002a), SROCC (Chen and Popovich, 2002b) and CCD (Székely et al., 2007) (defined in Section 2.3.1). iFAS also displays pixel-wise image differences with the purpose of providing local information about the fidelity measure behavior. The pixel-wise image difference has associated a histogram. This histogram can be considered to select the appropriated statistics for computing the global fidelity index during a fidelity measure design process. iFAS possesses box plots and multiple statistical comparisons between a set of fidelity assessment measures with the purpose of identifying if there are statistically significant differences in terms of the performance between them. iFAS also simulates model building by using regression analysis using training and test subsets. In Appendix B we describe in detail this software tool as well as we show how to use it in the fidelity assessment tasks from Chapters 5 and 6.

2.5 Conclusions

In this Chapter, we have studied the background information concerning image fidelity assessment. We have differentiated between image fidelity assessment and image quality assessment. Image fidelity assessment refers to the ability to quantify visual differences between a reference and test sample. Image quality assessment refers to the preference for one image over another. Afterwards, we have studied the subjective evaluation of visual differences typically perceived and reported by human observers.

In general, the standard way of measuring visual differences is based on subjective testing performed by human observers. In subjective testing, a set of well selected images/videos samples is shown to human observers with the purpose of collecting their opinion over the test samples following well-defined rules. The result of subjective tests is typically a database, termed image fidelity/quality database, intended for evaluating the performance of numerical fidelity measures. The performance of fidelity assessment measures is evaluated by comparing the numerical fidelity values with the scores reported by the human observers. Particularly, the following indices are typically reported in fidelity assessment reports and scientific papers PCC, SROCC, CCD, RMSE and MAE.

We developed an open source software tool designed to assist researchers, engineers and other users in the process of image fidelity assessment, named image Fidelity Assessment (iFAS). iFAS provides the following basic image fidelity assessment tools: computation of fidelity measures on a single pair of images and/or in a full database, visualization of pixel-wise image differences and histogram of the image differences, scatter plots and correlation analysis between human scores and objective measures. The correlation analysis is performed based on the most recent tools for the process of image fidelity assessment evaluation such as global correlation comparison, pairwise comparisons of correlations per reference, regression analysis and model building.

The contributions reported in this Chapter resulted in one international conference proceedings (Ortiz-Jaramillo et al., 2017) and one image fidelity software tool (Ortiz-Jaramillo, 2017b).

3

Objective quality estimation of compressed video sequences

3.1 Introduction

The most well-known fidelity assessment task is to determine how close a compressed image/video is to a given reference (distortion free or "perfect" video), i.e., the objective estimation of quality of compressed image/video sequences. In this Chapter we use the term quality instead of fidelity solely for the purpose of agreeing with the state-of-the-art terminology. However, in the rest of this thesis we clearly differentiate between both terms.

Today, due to the massive amounts of video produced, transmitted and stored in surveillance, broadcasting and other applications, video compression has become essential. Unfortunately, video compression tends to go hand in hand with reduced video quality. Therefore, quality assessment (quantification/measuring) and quality control (maintaining required quality levels) are very important tasks for increasing the user satisfaction. Since the end-user is often a human observer, quality control should include measures that mimic the *user-perceived video quality* (PVQ) (ITU, 1996). Thus, quality assessment of compressed videos has an important role in evaluating and improving the performance of today's video systems.

Methods for video quality assessment can be grouped into two categories: subjective and objective assessment (Liu et al., 2013b). Subjective assessment is typically performed by a group of humans, who evaluate videos according to certain well-defined criteria (see Chapter 2 for details) such as those defined in the related ITU standards (ITU, 1998). Often, the result of such an assessment is a Mean Opinion Score (MOS) or a Difference-MOS (DMOS) per assessed video sequence. Although MOS and DMOS do not fully characterize the response of human subjects (e.g. no information about the rating scale, about the variability of the human ratings (Winkler, 2009; Hobfeld et al., 2011)),

these measures are considered the most important parameters in characterizing subjective rating of video-based systems (ITU, 1998; Video-Quality-Experts-Group, 2003; Winkler, 2009). Additionally, when a sufficiently large group of human subjects is available, this method is the most well-known and most widely used technique for measuring PVQ of video-based systems (ITU, 1998; Video-Quality-Experts-Group, 2003). However, such a technique is in general complex, expensive, and time consuming. Therefore, it is unpractical for real time video processing and hard to incorporate into a system design process (Liu et al., 2013b). For this reason many researchers have proposed objective (numerical) methods for predicting PVQ directly from the video data, termed *video quality metrics* (VQMs). Currently, there exists a large variety of objective methods, ranging from simple ones employing local spatio-temporal statistics, loss of detail, and additive impairments, to more complex ones, such as those based on the results of physiological and/or psychovisual experiments (Pinson and Wolf, 2004a; Wang and Li, 2007; Seshadrinathan and Bovik, 2010; Li et al., 2011b; Liu et al., 2013b; Ortiz-Jaramillo et al., 2014a).

However, these objective methods are computationally too complex and/or not generic enough for a wide variety of video content scenes. The latter problem is mainly due to the strong dependency of VQMs on the video content (Feghali et al., 2007; Le-Callet et al., 2007; Khan et al., 2009; Huynh-Thu and Ghanbari, 2008; Korhonen and You, 2010; Garcia et al., 2010a; Pitrey et al., 2012; Ou et al., 2014). Despite this dependency being well-known, only few existing quality measures directly account for the effects of content. For instance, (Feghali et al., 2007; Garcia et al., 2010a; Korhonen and You, 2010; Ou et al., 2014) proposed models that combine *content related indices*, peak signal to noise ratio (PSNR), bit rate, spatial and temporal resolution for estimating the quality of compressed video sequences. Although those methodologies were tested only on few typical test videos showing no generalization power, they have shown that incorporating content in the VQM computation considerably improves the correlation between subjective and objective quality assessment. More important, these type of VQMs have shown to be of low computational complexity (Feghali et al., 2007; Khan et al., 2009; Korhonen and You, 2010; Garcia et al., 2010a; Ou et al., 2014).

Another major issue concerning objective video quality assessment of compressed sequences is the limited evaluation of the state-of-the-art VQMs. Typically, the methods are tested on databases including few testing samples (source video sequences), exhibiting little variation in the scene content (e.g. (Winkler, 2010) has concluded that, overall, a public video quality database covers about $10 - 20\%$ of the possible range in the spatial and temporal information axes), spatial/temporal resolution, and/or not being publicly available. For instance, Table 3.1 summarizes the number of source video sequences used to evaluate VQMs. Note that the maximum number of source sequences used for evaluating video quality algorithms is 59 which is one of biggest subjective video quality tests performed in the field (VQEG, 2010). In the most recent review concerning objective video quality assessment presented in (Chikkerur et al.,

Table 3.1: Number of source video sequences (N) used to evaluate VQMs

Reference	N	Database name
(Li et al., 2011b)	10	`LIVE` (Seshadrinathan et al., 2010)
(Pinson and Wolf, 2004a)	20	`VQEG-FR` (vqe, 2000)
(Wang and Li, 2007)	20	`VQEG-FR` (vqe, 2000)
(Seshadrinathan and Bovik, 2010)	20	`VQEG-FR` (vqe, 2000)
(Ortiz-Jaramillo et al., 2014a)	20	(10) `LIVE` (Seshadrinathan et al., 2010) and (10) `IVP` (Zhang et al., 2011)
(Moorthy and Bovik, 2010)	30	(10) `LIVE` (Seshadrinathan et al., 2010) and (20) `VQEG-FR` (vqe, 2000)
(Chikkerur et al., 2011)	30	(10) `LIVE` (Seshadrinathan et al., 2010) and (20) `VQEG-FR` (vqe, 2000)
(VQEG, 2010)	59	(59) `VQEGHD` (VQEG, 2010)

2011), only two public databases have been used for comparing performance of the considered VQMs. In general, this is too little data for drawing conclusions about the VQMs.

In this thesis, we aim to substantially improve on this issue by performing the evaluation on a total of 696 distorted sequences (corresponding to a total of 102 source sequences covering about $50-60\%$ of the possible range in the spatial and temporal information axes, cf. Figure A.10), differing in content as well as in temporal and spatial resolution.

Another important contribution of this thesis is the proposition of a novel method for objective video quality assessment of compressed sequences. The proposed method involves off-line training of the mapping functions' parameters and their relationship to video content characteristics. First, the off-line training of parameters is done by computing a VQM between a reference (source) sequence and several compressed versions of that sequence. Next, assuming the (D)MOS values are known for multiple levels of distortion, we tune a mapping function to predict (D)MOS from the VQM, i.e., the parameters of the mapping function are tuned specifically to each source content. Finally, we model the relationship between the parameters of the mapping function and the video content characteristics (extent of image details and motion of the video sequence). Also, we perform an extensive experimental evaluation based on a total of four VQMs (the structural similarity index measure (Wang et al., 2004), the standardized method for objectively measuring video quality (Pinson and Wolf, 2004a), the video quality measure based on decoupling detail losses and additive impairments (Li et al., 2011b) and the peak signal to noise ratio [PSNR]), each tested on 696 distorted video sequences. For the considered VQMs, we explore 105 *content related indices* to model the relationship between the VQM and the DMOS. The 696 test video samples (102 source video sequences) were taken from five public databases (`IRCCyN IVC Influence Content` (Pitrey et al., 2012), `CIF` as well as `4CIF EPFL-PoliMI` (De-Simone et al., 2009), `IRCCyN IVC 1080i` (Pechard et al., 2011) and `IVP` (Zhang et al., 2011)). Additionally, we provide guidelines for using the proposed approach to select an appropriate set of video distortion levels for the purpose of subjective

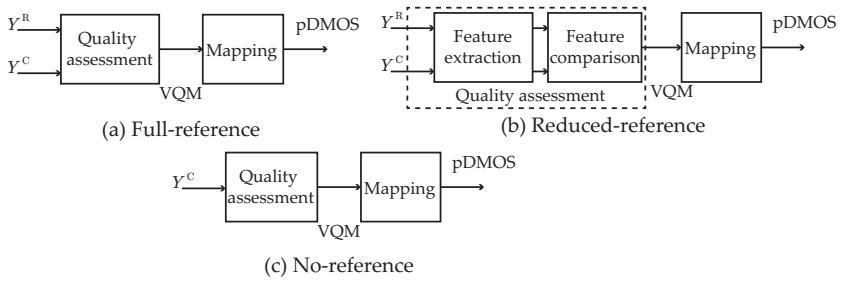

Figure 3.1: General framework for video or image quality assessment depending on the reference availability. (a) Full-reference, (b) reduced-reference and (c) no-reference framework.

quality assessment studies of compressed video sequences.

Our experimental results suggest that when adequately combined with content related indices, even very simple distortion measures (such as PSNR) are able to achieve high performance, i.e., high correlation between the VQM and the PVQ. Specifically, we have found that by incorporating video content features, it is possible to increase the performance of a VQM by up to 20% relative to its non-content-aware baseline.

This Chapter is organized as follows. In Section 3.2, current approaches dealing with objective evaluation of quality of compressed video sequences are discussed. Afterwards, we explore multiple factors affecting the relationship between VQMs and PVQ under different source sequence. Later, Section 3.3 discusses the proposed method and its implementation details. Thereafter, in Section 3.4, we present and discuss the results obtained in our tested data. Finally, in Section 3.5, we draw conclusions and propose future work.

3.2 Background

In the following, we provide a summary of the state-of-the-art of objective VQMs and describe the effects of video content on some of the most well-known and most widely used objective VQMs.

3.2.1 Objective video quality measures

Objective VQMs use computer algorithms for computing numerical scores on corrupted video sequences that should agree with the subjective assessment provided by human evaluators. In general, VQMs are categorized as full-reference, reduced-reference or no-reference, depending on the availability of a reference (Chikkerur et al., 2011). In Figure 3.1, Y_R and Y_C are the reference and corrupted video sequences, respectively. In either case, the final predicted quality value, termed predicted (D)MOS [p(D)MOS], is typically obtained by

applying a predefined mapping to the quality measure (Video-Quality-Experts-Group, 2003). Note that fidelity assessment covers only the cases shown in Figures 3.1 (a) and (b) because in fidelity assessment, the reference or some of its features are known (see Chapter 2). Therefore, we do not further explore algorithms following the framework of Figure 3.1 (c).

Algorithms following the approach of Figures 3.1 (a) and (b) can be further classified into traditional point-based methods, image characteristics based methods or perceptual oriented methods, depending on the set of techniques used to compute the quality measure (Chikkerur et al., 2011). Traditional point-based methods (TPB) use pixel-wise operations for computing differences between images and/or video sequences. For instance, PSNR is the most simple but still widely used traditional point-based method (Liu et al., 2013b). Image characteristics based methods use statistical measures (mean, variance, histograms) in local neighbourhoods and/or visual features (blur estimates, block distortion measures, texture characteristics) for computing numerical scores. For example, the standardized method for objectively measuring video quality (SOVQM) (Pinson and Wolf, 2004a) is computed by using local spatio-temporal statistics which are computed on blocks of a fixed size. Afterwards, the extracted features from reference and corrupted sequences are thresholded, compared and pooled to obtain an unique numerical quality measure.

Another example from this category is the well-known structural similarity index (SSIM) (Wang et al., 2004) which uses statistics (mean and standard deviation) of neighbouring pixels to characterize luminance, contrast and structure of the reference and corrupted sequences. Thereafter, features of the reference and corrupted sequences are compared and pooled obtaining a numerical quality measure. Another technique using image characteristics is the video quality measure based on decoupling detail losses and additive impairments (VQAD) (Li et al., 2011b) which subtracts a restored version of the corrupted sequence from the reference sequence. This subtraction is made to differentiate between distortions due to detail losses (edges, high textured regions and/or small objects) and distortions due to introduced impairments such as blocking artifacts, noise and/or false edges. Afterwards, the detail losses and introduced impairments higher than a threshold are individually pooled and linearly combined to predict the quality of the corrupted sequence.

Perceptual oriented methods have been designed based on the results of physiological and/or psychovisual experiments. This approach includes among others, modelling human visual attention and modelling human speed perception (Wang and Li, 2007; Seshadrinathan and Bovik, 2010; Ortiz-Jaramillo et al., 2014a). For instance, the weighted structural similarity index (wSSIM) (Wang and Li, 2007) uses the structural similarity index for measuring local image similarities, termed quality maps. For computing a unique quality score from those quality maps, a spatiotemporal weighted mean based on saliency maps is used. The saliency map is computed based on a statistical model of speed perception derived from psychovisual experiments conducted in (Stocker and Simoncelli, 2006). The weighted temporal quality met-

ric (wTQM) (Ortiz-Jaramillo et al., 2014a) computes temporal distortions directly from optical flows and models the human visual attention using saliency maps on the pooling strategy. Such saliency maps were computed based on the results of psychovisual experiments conducted by the authors (Ortiz-Jaramillo et al., 2014a). The motion-based video integrity evaluation index (MOVIE) (Seshadrinathan and Bovik, 2010) uses a Gabor filter bank specifically designed based on physiological findings for mimicking the visual system response. The video quality evaluation is carried out using two components (spatial and temporal distortions). The spatial distortions are computed as squared differences between Gabor coefficients and the temporal distortions are obtained from the mean squared error between reference and corrupted sequences along motion trajectories (Seshadrinathan and Bovik, 2010). Thereafter, both distortions are combined to predict the quality of the corrupted sequence. Noteworthy is that the above methods do not account directly for content information and instead use mechanisms to mimic the visual system under certain conditions. Additionally, these methods are often computational complex (Seshadrinathan and Bovik, 2010; Ortiz-Jaramillo et al., 2014a).

More recent studies in the video quality assessment field include saliency map estimation and machine learning algorithms. The saliency maps are typically used as weights to compute a weighted average of pixel wise differences (Zhang and Liu, 2017; Radun et al., 2017; Garcia-Freitas et al., 2018). Machine learning algorithms are normally used to combine a set of predefined features with the purpose of estimating quality differences. For instance, in (Menor et al., 2016) the following parameters where estimated and combined using a neural network to estimate video quality of compressed video sequences: jerkiness, blur, blockiness, luminance distortion, chrominance distortions and temporal distortions. In (Torres-Vega et al., 2017), deep learning has been used for the quality assessment of live video streaming. A machine learning technique called extreme learning machine is employed in (Wang et al., 2016) in the pooling strategy. In (Kim and Lee, 2017), a convolutional neural network is used to predict the visual weight of each pixel based on the error map.

To the best of our knowledge, only few methods in the literature explicitly use content information for video quality assessment. For instance, (Feghali et al., 2007) use PSNR, frame rate and average motion magnitude to estimate quality of low resolution video sequences. In contrast to our proposed method, (Feghali et al., 2007) do not take into account the saturation effect of human vision which is better modeled by using a S-shape function (Huynh-Thu and Ghanbari, 2008; Ou et al., 2011). Another disadvantage of their approach is that PVQ is affected by both spatial and temporal characteristics of the video sequence (Le-Callet et al., 2007) and the average motion magnitude is not enough for accounting video content characteristics. (Khan et al., 2009) acknowledged the importance of *content related indices* as they investigated the impact of packet loss on video by identifying minimum quality requirements of the system under specific video content. However, that work does not specifically propose a VQM. (Garcia et al., 2010a) used *content related indices*

extracted from the encoded data (block based motion vectors, discrete cosine transform coefficients, number of macro blocks per frame) and the bit-rate for modeling the quality of high definition compressed video sequences but, unlike the proposed method, the method proposed in (Garcia et al., 2010a) needs prior information about the testing data which in general is not available.

(Rodriguez et al., 2014) investigated the impact of video content preference in measuring the quality of video streaming applications. The method uses a non-linear combination of the following technical parameters as quality index: number, duration and temporal location of pauses that occur during a video streaming transmission. Additionally, a so-called content preference function is used to adjust the quality index value. Unlike the proposed method, the scheme of (Rodriguez et al., 2014) needs to store video content preferences and classify beforehand each source sequence into one of the content-type categories defined by the authors (sports, news, or documentary) which in general is unpractical in the design of VQMs (Garcia et al., 2010a). Recently, (Ou et al., 2014) proposed a measure for estimating the quality of compressed video sequences by using quantization step of the coder, normalized motion activity, standard deviation of frame differences, Gabor features, spatial and temporal resolution. The measure estimates 3 different mapping function parameters using the linear combination of *content related indices*. Although their measure is similar to the method considered in our work, it has several comparative drawbacks. For instance, the measure of (Ou et al., 2014) uses what the authors call normalized MOS which depends on the perceived quality of the video sequence under maximum spatial resolution, maximum frame rate and minimum quantization level, which in general are not available for typical video-based applications. Also, the measure of (Ou et al., 2014) is highly dependent on the range of spatial resolutions and frame rates used during the training phase.

Our method is based on the work of (Korhonen and You, 2010) who propose a measure that combines the standard deviation of Sobel filtered images, the standard deviation of frame differences and PSNR for estimating the quality of three source sequences from the `CIF EPFL-PoliMI Video Quality Assessment Database` (De-Simone et al., 2009). In that work an exponential function was used as mapping function which, unlike our proposed method, does not take into account the saturation effect of the human vision. In addition, (Korhonen and You, 2010) test linear models of one independent variable from a set of four *content related indices*. By contrast, we test linear combinations of two independent variables from a set of 105 selected *content related indices* including both spatial and temporal information axes in the parameter estimation agreeing with the fact that PVQ is affected by both types of *content related indices* (Le-Callet et al., 2007). Finally, compared to the measure of (Korhonen and You, 2010) which used only three source sequences for testing, our study presents a stronger statistical analysis based on the results obtained on five different video quality databases (102 source sequences).

3.2.2 Effects of video content on video quality measures

In this thesis, we explore the relationship between PVQ and VQMs (SSIM (Wang et al., 2004), SOVQM (Pinson and Wolf, 2004a), VQAD (Li et al., 2011b) and PSNR) under varying content (different source sequence). First we explore the most appropriated mapping function VQM→DMOS by considering the following twelve linear and nonlinear monotonically decreasing/increasing functions: (a) linear, (b) quadratic, (c) cubic, (d) exponential, (e) logistic, (f) hyperbolic, (g) cosine, (h) logarithmic, (i) rational, (j) complementary error, (k) complementary cumulative raised cosine, (l) complementary cumulative log-laplace. Previous listed functions were selected based on inspection of the experimental data computed on the `IRCCyN IVC 1080i` video quality database (Pechard et al., 2011). We select from the set of listed functions, the function with the best fit to the data by means of statistical analysis. Specifically, the selection was performed by using multiple statistical comparisons as discussed in (Garcia et al., 2010b) and Section 2.3. The objective of this test is to determine if we may conclude from the data that there are differences in terms of performance among the tested functions. The performance is measured using the correlation indices described in Chapter 2. From the multiple statistical comparisons we found that the best performing functions are:

- for PSNR: (b) quadratic, (c) cubic, (i) rational, (j) complementary error and (k) complementary cumulative raised cosine (correlations higher than 92%);
- for SSIM: (b) quadratic, (c) cubic, (j) complementary error and (k) complementary cumulative raised cosine (correlations higher than 90%);
- for SOVQM: (b) quadratic, (c) cubic, (e) logistic, (j) complementary error and (k) complementary cumulative raised cosine (correlations higher than 93%);
- for VQAD: (a) linear, (b) quadratic, (c) cubic, (e) rational, (j) complementary error and (k) complementary cumulative raised cosine (correlations higher than 90%).

The listed functions above are the best performing functions tested in the different VQMs, i.e., there are no significant differences between them (p-values higher than 0.1) but they perform significant better than the other tested functions (p-values lower than 0.05). Although (a) linear, (b) quadratic, (c) cubic and (i) rational perform well, they do not account for the saturation effect of the human vision which is a very important effect when measuring PVQ. That is, the human vision has little sensitivity to small changes in quality in the ranges of very low or very high levels of image quality (Haakma et al., 2005). Therefore, a S-shape function is more desirable, e.g., (e) logistic, (j) complementary error or (k) complementary cumulative raised cosine, to take into account the saturation effect of human vision. Note that some public video quality databases includes only 4 distortion levels per scene (e.g., `IRCCyN`

IVC Influence Content (Pitrey et al., 2012)), which greatly limits the number of data points available in the training process. That is, due to limitations in the current available data, it is inconvenient to model the relationship VQM→DMOS with more than 2 parameters. Therefore, it is important to keep the number of parameters limited for avoiding over fitting and poor generalization power of the trained models.

Therefore, following the above multiple statistical comparisons and taking into account the saturation effect of the PVQ as well as the limitations due to number of data points, we choose the complementary error function for the 4 tested VQMs. The complementary error function is defined as

$$f(x;\mathbf{a}) = 1 - \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-a_1}{a_2\sqrt{2}}\right)\right), \tag{3.1}$$

where $\mathbf{a} = [a_1, a_2]^T$ is a vector of parameters with the best fit to the data. a_1 controls the x-axis bias and a_2 the slope of the function. Here, erf is defined as follows

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-w^2\right) dw.$$

The parameters of this function can be related to the saturation effect of human vision and the rate of change between the VQM and the PVQ which are the most affected parameters under different source video (Korhonen and You, 2010). On the one hand, the rate of change controls how fast the VQM should drop or rise depending on the content, i.e., it controls the rate between DMOS/VQM. From the PVQ point of view, it is the minimum change in the VQM to get a perceived quality difference. For instance, for a high textured sequence (natural scenes) the rate of change should be smaller than for a low textured sequence (cartoon scenes) because distortions are easier to perceive in the former type of scene. On the other hand, the saturation effect of human vision is controlled by using the called halfway point of the S-shape curve. That is, the VQM value in which (D)MOS equals to 0.5 (in a 0-1 (D)MOS range, see Figure 3.2). From the PVQ point of view, it controls the saturation point of quality, i.e., it controls where, in the VQM axis, the human vision has higher sensitivity to small changes in quality.

(Huynh-Thu and Ghanbari, 2008) have studied experimentally the reach of PSNR as VQM. The authors found that PSNR is a good indicator of quality when the content and distortion type are fixed. For instance, Figure 3.2 shows the plot of DMOS in function of PSNR for the six cases of source sequences from the CIF EPFL-PoliMI Video Quality Assessment Database (see Appendix A.5 for detailed description of the tested sequences) (De-Simone et al., 2009). The solid lines represent curves with the best fit to the data, i.e., pDMOS correlates well with the DMOS. Each regression line (mapping function) was obtained by using only the test sequences corresponding to the same source sequence. That is, the points marked with the same symbol. The black line represents a curve with the best fit to the whole set of data points. The PCC between the DMOS and the data represented by the black line is 0.74, which is low for such a

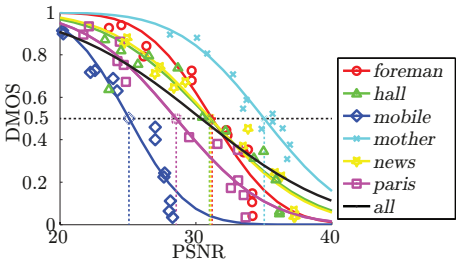

Figure 3.2: DMOS in function of PSNR. Each marker symbol represents a different source sequence. The solid lines represent curves with the best fit to the data. The dotted lines are halfway points, i.e., PSNR values such that pDMOS= 0.5.

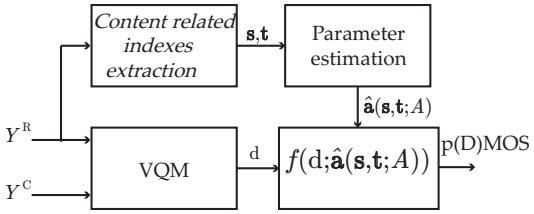

Figure 3.3: Framework of the proposed method. d and $\hat{\mathbf{a}}$ denote, respectively, the numerical value of the quality measure and the estimated parameters for the mapping function. $\hat{\mathbf{a}}$ is estimated by using the matrix A (obtained during the off-line training) and SA ($\mathbf{s}$) as well as TA ($\mathbf{t}$).

small database (Keimel et al., 2009). If the regression line or mapping function parameters are adapted to the current source sequence (non-black lines), the PCC increases by 30%, i.e., PCC = 0.96. These results suggest that there is a unique mapping function PSNR→DMOS when the reference content is fixed, i.e., the parameters of the mapping function depend mostly on the video content (Keimel et al., 2009). Also, (Keimel et al., 2009) stated that "*even simple measures can perform well when tuned to a specific source sequence*". However in practical applications the parameters of the mapping function are unknown a-priori. The challenge is therefore to find a method to adjust such parameters automatically to the current spatio-temporal video content at hand, i.e., the spatial activity and temporal activity of the source sequence.

3.3 Proposed method

Figure 3.3 shows the proposed framework for video quality assessment of compressed sequences including content information. In the quality measure step, a numerical VQM is computed on the reference (Y^{R}) and compressed (Y^{C}) video sequences obtaining a numerical value d. We extract the *content related*

indices from the reference sequence. Elements of vectors $\mathbf{s}$ and $\mathbf{t}$ are *content related indices* representing the spatial activity (SA) and temporal activity (TA) of the video sequence, respectively. The SA and TA are used as input to the parameter estimation block which is a simple linear model trained off-line using a set of training samples (see Section 3.3.2 for implementation details). Afterwards, the estimated parameters and the VQM value are used as input to the mapping function to estimate the quality of the corrupted sequence.

3.3.1 Off-line training for the proposed method

The purpose of the off-line training process is to estimate the coefficients of a matrix A which is a parameter of the mapping function f (see Figure 3.3). The training is performed using I source sequences and their J corrupted (distorted) versions, thus a total of $I \times (J+1)$ video sequences, for which the perceived quality scores ((D)MOS$_{i,j}$) are known. There, the matrix A describes the relationship between the following three components: (1) the values of the selected VQM computed for the training sequences, (2) the content-related indices of the training sequences, and (3) the perceived quality of the training video sequences ((D)MOS).

The off-line training starts by computing VQM values ($d_{i,j}$) between a reference sequence Y_i^{R} and its corrupted versions $Y_{i,j}^{\mathrm{C}}$ $\forall j = 1, \ldots, \mathrm{J}$, where J is the total number of available corrupted sequences of the ith source sequence. Thereafter, a non-linear regression method (in our case, the least absolute residual method (Bloomfield and Steiger, 1980)) is applied between VQM values and the corresponding available set of (D)MOS values for the ith source sequence. The result of the non-linear regression is the set of parameters ($\mathbf{a}_i$) for the mapping functions tuned specifically on the ith source sequences (see Figure 3.4(a)). $\mathbf{s}_i$ and $\mathbf{t}_i$ are computed for the ith source sequence with the purpose of characterizing the content information of the sequence (see Section 3.3.2 for details about the *content related indices* $\mathbf{s}_i$ and $\mathbf{t}_i$). Afterwards, the *content related indices* and the mapping function parameters ($\mathbf{a}_i$) are used to find a function $\Theta(\cdot)$ such that

$$\sum_{i=1}^{\mathrm{I}} \|\mathbf{a}_i - \Theta(\mathbf{s}_i, \mathbf{t}_i)\| \approx 0. \tag{3.2}$$

Therefore, the problem is to select such a function that models the relationship between the *content related indices* ($\mathbf{s}_i$, $\mathbf{t}_i$) and the mapping function parameters $\mathbf{a}_i$. In this thesis, we use a linear model for simplicity and we leave the model selection to future research. We will see later in Section 3.4 that higher order models or different model estimation methodologies (e.g., support vector regression) could lead to better results.

By assuming a linear model in Equation (3.2), we have the following simpli-

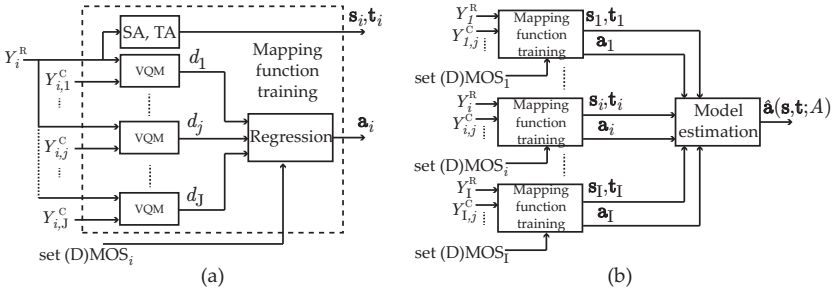

Figure 3.4: Flow chart of the off-line training for the proposed method. (a) Off-line training of the mapping function parameters $\mathbf{a}_i$ for the ith training source sequence using J corrupted sequences of the same source and their (D)MOS values (set (D)MOS$_i$). (b) Off-line identification of the model to estimate A by using *content related indices* $\mathbf{s}_i$ and $\mathbf{t}_i$ using I different training source sequences.

fication of the model estimation which can be solved by using linear regression.

$$\sum_{i=1}^{\mathrm{I}} \left\| \mathbf{a}_i - A \begin{bmatrix} 1 \\ \mathbf{s}_i \\ \mathbf{t}_i \end{bmatrix} \right\| \approx 0, \tag{3.3}$$

where I is the number of available training source sequences (see Figure 3.4(b)). Note that the performance of the method depends highly in the variety of video content included during the training phase. For instance, if we train our methodology using only high textured sequences (wild nature documentaries), we expect that the unknown incoming sequences possess similar content related characteristics to guarantee a similar performance compared with the training phase. That is, we anticipate in this example a lower performance for low textured sequences (cartoon sequences) than for high textured sequences.

In any case, after finding the matrix A during the off-line training, the model is ready for evaluating an arbitrary unknown incoming sequence by applying the following steps: (i) compute d value, $\mathbf{s}$ and $\mathbf{t}$, (ii) compute $\hat{\mathbf{a}} = A[1, \mathbf{s}, \mathbf{t}]^T$, and, (iii) predict the (D)MOS by mapping the obtained d value using the mapping function and the estimated $\hat{\mathbf{a}}$ parameters, i.e., $f(d; \hat{\mathbf{a}})$.

3.3.2 Implementation details

The implementation of the VQMs used in this work were obtained from the web pages of the authors (SSIM (Zhou et al., 2014), VQM (Pinson and Wolf, 2004b), VQAD (Li et al., 2011a)), except for the PSNR which was computed as

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{L}^2}{\mathrm{MSE}} \right),$$

where MSE is the mean squared error between the pixel intensities, i.e.,

$$\mathrm{MSE} = \frac{1}{\mathrm{NMK}} \sum_{n,m,k} \left(Y^{\mathrm{R}}(n,m,k) - Y^{\mathrm{C}}(n,m,k) \right)^2$$

for K frames of size $\mathrm{N} \times \mathrm{M}$. L is the maximum intensity value of Y^{R}. Note that for PSNR as well as for SSIM the quality increases when the VQM increases while for SOVQM as well as for VQAD the quality decreases when the VQM increases. Therefore, we use $f(x; \mathbf{a})$ for PSNR and SSIM and $1 - f(x; \mathbf{a})$ for SOVQM and VQAD.

Currently we have explored *content related indices* extracted from

- pixel-wise differences (magnitude of spatial and temporal gradients),
- spatial dependencies of pixel values (Gray level co-occurrence matrix GLCM (Randen and Husoy, 1999)),
- magnitude of optical flows (Lucas-Kanade algorithm (Barron et al., 1992)),
- the magnitude of spatial Sobel filtered images and
- the magnitude of SI13 filtered images (SI13 filter is a spatial filter designed specifically to measure perceptually significant edges by using a 13 pixels filter (Pinson and Wolf, 2004a)).

In particular, the following statistics were extracted as *content related indices*: energy, entropy, contrast, homogeneity as well as correlation per frame computed from the GLCM (Randen and Husoy, 1999), descriptive statistics per frame computed from

- the pixel-wise differences,
- the magnitude of optical flows,
- the magnitude of Sobel filtered images and
- the magnitude of SI13 filtered images.

The descriptive statistics are mean, median, standard deviation, skewness, kurtosis and total variation (sum of absolute values). Thereafter, the mean, the standard deviation and the maximum of those descriptive statistics per frame are computed as global *content related indices*.

That is, 15 *content related indices* on GLCM ({energy, entropy, contrast, homogeneity, correlation}$\times$ {mean, standard deviation, maximum} = 15) and 18 *content related indices* on 5 spatial and temporal features ({mean, median, standard deviation, skewness, kurtosis, total variation} $\times$ {mean, standard deviation, maximum} $\times$ {spatial pixel-wise differences, temporal pixel-wise differences, the magnitude of optical flows, the magnitude of Sobel filtered images,

the magnitude of SI13 filtered images$\} = 18 \times 5 = 90$), resulting in a total of $15 + 90 = 105$ *content related indices*.

Before showing the best performing linear models, we describe the individual used *content related indices*:

- s_1 is the mean value of the magnitude of the SI13 image:

$$s_1 = \frac{1}{NMK} \sum_{n,m,k} |\mathrm{SI13}\{Y^{\mathrm{R}}\}(n,m,k)|,$$

  where $|\mathrm{SI13}\{Y^{\mathrm{R}}\}(n,m,k)|$ is the magnitude of Y^{R} filtered by using the SI13 filter in the (m,n)th pixel of the kth frame (Pinson and Wolf, 2004a).

- s_2 is the mean skewness over all frames of the magnitude of the SI13 image:

$$s_2 = \frac{1}{K} \sum_{k} \frac{\frac{1}{NM-1} \sum_{n,m} \left(|\mathrm{SI13}\{Y^{\mathrm{R}}\}(n,m,k)| - s_1 \right)^3}{\left(\frac{1}{NM} \sum_{n,m} \left(|\mathrm{SI13}\{Y^{\mathrm{R}}\}(n,m,k)| - s_1 \right)^2 \right)^{3/2}}.$$

- s_3 is the mean contrast over all frames of the gray level co-occurrence matrix:

$$s_3 = \frac{1}{K} \sum_{k} \sum_{x,y} C(x,y,k) \log \left(C(x,y,k) \right),$$

  with $C(x,y,k)$ representing a count of the number of times that $Y^{\mathrm{R}}(n,m,k) = x$ and $Y^{\mathrm{R}}(n+\Delta n, m+\Delta m, k) = y$ in the kth frame, where $(\Delta n, \Delta m) \in \{(0,1),(-1,1),(-1,0),(-1,-1)\}$ (Randen and Husoy, 1999).

- t_1 is the mean total variation over all frames of the temporal gradient:

$$t_1 = \frac{1}{NMK} \sum_{n,m,k} \left| Y^{\mathrm{R}}(n,m,k) - Y^{\mathrm{R}}(n,m,k-1) \right|.$$

- t_2 is the maximum across all frames of the total variation of the temporal gradient:

$$t_2 = \max_{k} \sum_{n,m} \left| Y^{\mathrm{R}}(n,m,k) - Y^{\mathrm{R}}(n,m,k-1) \right|.$$

After introducing the individual *content related indices*, the best performing linear models are described in the following paragraphs. For the VQAD case, none of the tested linear combinations of *content related indices* performed well in modeling the parameters of the mapping functions VQAD→DMOS.

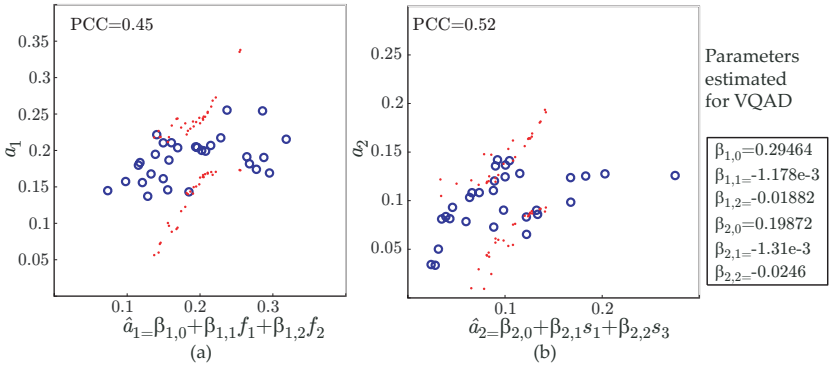

Figure 3.5: Scatter plot of (a) a_1 in function of $\hat{a}_1$ and (b) a_2 in function of $\hat{a}_2$ for VQAD to DMOS mapping functions. The dots are the confidence interval for $\hat{a}_1$ and $\hat{a}_2$. Note that the parameters in the plots were estimated using the databases `IRCCyN IVC 1080i` and `IVP` described in Appendix A.5

For instance, Figure 3.5(a) and (b) show the plot of the parameters of the mapping functions of VQAD measure (a_1 and a_2) versus the estimated parameters using content features ($\hat{a}_1$ and $\hat{a}_2$) where

$$f_1 = \max_k \sum_{x,y} C(x,y,k) \log\left(C(x,y,k)\right)$$

and f_2 is the mean across the time of the skewness computed on the temporal pixel-wise differences. Each circle represents the plot of $\mathbf{a}_i$ optimized for the ith source sequence versus the parameters estimated using the spatial and temporal *content related indices*. The model in the Figure is the best performing linear combination of *content related indices*. However, this model show a weak correlation between the estimated parameters and the tested indices (PCC <0.6). This may be due to the complexity of the VQAD measure which includes two different masking mechanisms (spatial and temporal masking) (Li et al., 2011b). This kind of mechanism to mimic the visual system makes it more difficult to identify the relationship between the VQM and DMOS under varying source sequence.

For the remaining VQMs, we have proposed the following model that combines spatial and temporal *content related indices* for computing the parameters of the mapping functions VQM→DMOS under study.

$$\mathbf{a} \approx A \begin{bmatrix} 1 \\ \mathbf{s} \\ \mathbf{t} \end{bmatrix} = \begin{bmatrix} \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} & \alpha_{1,5} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} & \alpha_{2,5} \end{bmatrix} \begin{bmatrix} 1 \\ s_1 \\ s_2 \\ s_3 \\ t_1 \\ t_2 \end{bmatrix}, \tag{3.4}$$

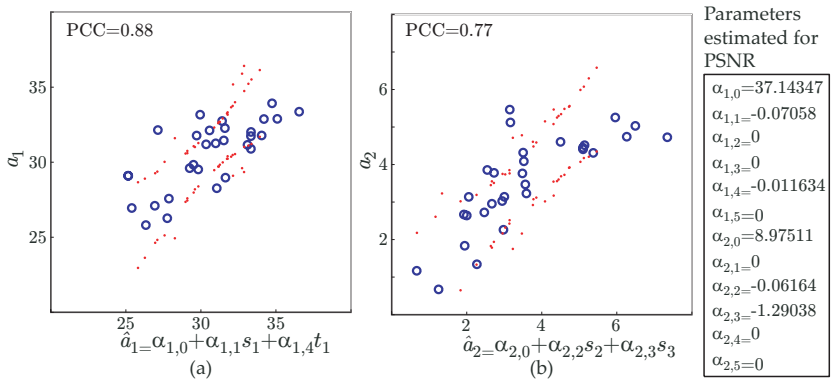

Figure 3.6: Scatter plot of (a) a_1 in function of $\hat{a}_1$ and (b) a_2 in function of $\hat{a}_2$ for PSNR to DMOS mapping functions. The dots are the confidence interval for $\hat{a}_1$ and $\hat{a}_2$. Note that the parameters in the plots were estimated using the databases `IRCCyN IVC 1080i` and `IVP` described in Appendix A.5

where $\alpha_{p,q}\forall\ p,q$ are estimated off-line for each VQM as explained in Section 3.3.1. Some of the parameters are set to zero depending on the VQM. For instance, based on experimental results, we found that among the *content related indices* tested in this thesis, the following are good predictors for the parameters of the mapping function PSNR→DMOS (Figure 3.6):

$$\hat{a}_1 = \alpha_{1,0} + \alpha_{1,1}s_1 + \alpha_{1,4}t_1$$

$$\hat{a}_2 = \alpha_{2,0} + \alpha_{2,2}s_2 + \alpha_{2,3}s_3.$$

For SSIM and SOVQM we have also explored different linear combination of *content related indices* and we found that the models in Figures 3.7 and 3.8 are the best performing models for the parameters of the mapping functions SSIM→DMOS and SOVQM→DMOS, respectively. The dots in Figures 3.5, 3.6, 3.7 and 3.8 are the confidence intervals computed for $\hat{a}_1$ and $\hat{a}_2$. That is, it is very likely that $\hat{a}_1$ and $\hat{a}_2$ lie within the confidence interval for an incoming test sample. This can be used as an indication of stability of the model. For instance, the model in Figure 3.6 is more stable and accurate than the models in Figures 3.5, 3.7 and 3.8 because the interval that contains the true value for $\hat{\mathbf{a}}$ is smaller for PSNR model than for SSIM, SOVQM and VQAD. That is, the prediction error between $\mathbf{a}$ and $\hat{\mathbf{a}}$ is smaller in PSNR model than SSIM, SOVQM and VQAD models. This can be shown as well with the respective PCC values also shown in the plots.

From the confidence intervals we can conclude that the proposed method is going to perform well using PSNR but not using the other tested VQMs. That is, the model using PSNR is the only one able to predict proper parameters for the mapping function in the tested samples. The other models are expected to

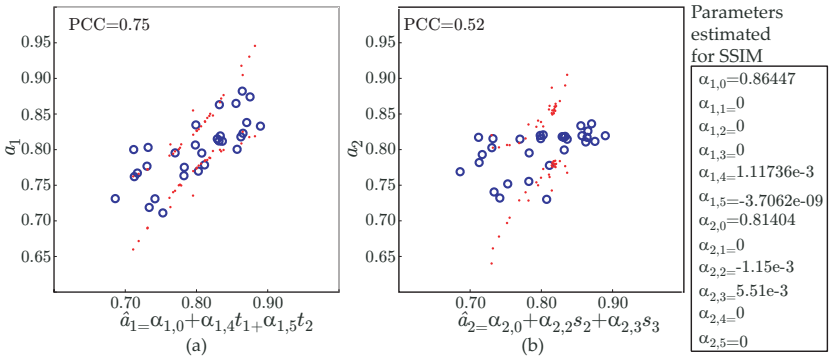

Figure 3.7: Scatter plot of (a) a_1 in function of $\hat{a}_1$ and (b) a_2 in function of $\hat{a}_2$ for SSIM to DMOS mapping functions. The dots are the confidence interval for $\hat{a}_1$ and $\hat{a}_2$. Note that the parameters in the plots were estimated using the databases `IRCCyN IVC 1080i` and `IVP` described in Appendix A.5

perform poorly because they predict parameters with very large errors degrading even the performance of the VQM as the results will show later. The poor stability of VQAD, SSIM and SOVQM models show a potential disadvantage of the proposed method because it means that there is not guarantee of finding a relationship VQM→DMOS under different source reference for particular measures (at least not with the *content related indices* tested in this work). Additionally, the trend of the data points for the parameter $\hat{a}_2$ shows that this parameter is better modeled by using an exponential like function than by a linear model. Nevertheless, we will show later in Section 3.4 that the proposed method has also major advantages when the VQM is a TPB method such as the PSNR.

3.4 Results and Discussion

In this Section, we present and discuss the obtained results. After that, in Section 3.4.2 we introduce a novel method for selecting the distorted videos for a subjective test of video quality such that their perceived quality is uniformly distributed over the whole quality range (e.g. measured DMOS values uniformly sample the range of 1 to 100).

3.4.1 Evaluation of the proposed method

We add the prefix letter C (standing for *content-aware*) to every VQM acronym with the purpose of differentiating between the performance of the original VQM and the same measure using our proposed method, e.g., PSNR is the original VQM and CPSNR is the quality prediction by using the VQM and the proposed method.

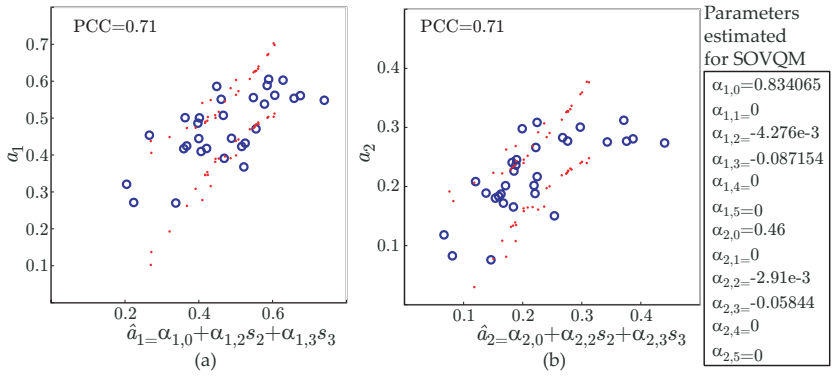

Figure 3.8: Scatter plot of (a) a_1 in function of $\hat{a}_1$ and (b) a_2 in function of $\hat{a}_2$ for SOVQM to DMOS mapping functions. The dots are the confidence interval for $\hat{a}_1$ and $\hat{a}_2$. Note that the parameters in the plots were estimated using the databases `IRCCyN IVC 1080i` and `IVP` explained in Appendix A.5

Figure 3.9 shows the performance of the considered VQMs discussed in Section 3.3.2. Databases `IRCCyN`, `IVP` are used for appraising the performance indices (PCC, SROCC, RMSE and MAE, see Section 2.3). Note that the performance for PSNR, SSIM, SOVQM and VQAD were computed after fitting the selected mapping function without using any content information, i.e., $\alpha_{i,j} = 0$ for $i = 1, 2$ and $j = 1, \ldots, 5$. For comparing the performance between the VQMs, we use cross validation with a repeated random sub-sampling procedure using 100 iterations as discussed in (Witten et al., 2011). At every iteration, the total number of 30 source contents on `IRCCyN` and `IVP` databases is randomly split into two mutually exclusive sets, termed training and validation sets. The coefficients of the matrix A are estimated with the training set (18 sequences) and the accuracy is assessed by using the validation set (12 sequences). The partition is made such that the training phase has always 6 source sequences from `IVP` database and 12 source sequences from `IRCCyN` database with the purpose of producing equal proportions of each database.

Scatter plots (a), (b) and (c) in Figure 3.9 show the PCC and the SROCC as well as their confidence intervals computed for the considered sets of the test sequences where the value of 1 indicates high correlation and 0 is no correlation between the tested quality measure and the DMOS. In the scatter plots, the closer the data points to the top right corner, the better the VQM performance. For instance, the best performing methods according to the plots are SOVQM followed by CPSNR (PSNR using the proposed method), CSOVQM and VQAD. Noteworthy is that the performance of PSNR increases from 0.68 to 0.80, i.e., about 17% (30% in linear Fisher's Z) by the proposed method, confirming the power of the proposed approach in TPB methods (the increase was computed as explained in Chapter 2).

There is an increase from 0.67 to 0.72 in the correlation between DMOS

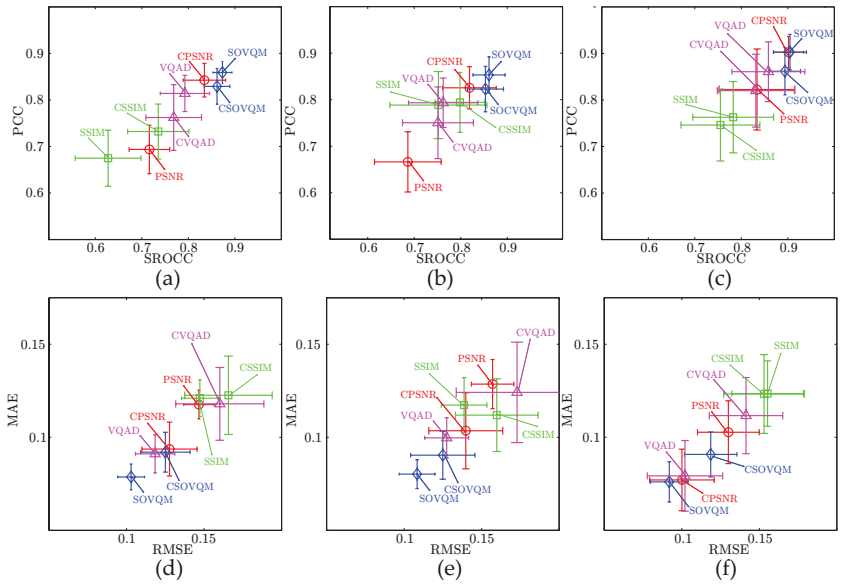

Figure 3.9: Performance of the considered video quality measures appraised on `IRCCyN` and `IVP` databases. The proposed method is named CPSNR, CSSIM, CSOVQM and CVQAD (we add the prefix letter C to every VQM acronym). Scatter plots of PCC and SROCC for (a) All data, (b) `IRCCyN` and (c) `IVP`. Scatter plot of MAE and RMSE for (d) All data, (e) `IRCCyN` and (f) `IVP`.

and CSSIM compared to SSIM, i.e., about 7% (12% in linear Fisher's Z). Even though the model in Figure 3.7(b) does not predict accurately the a_2 parameter, the proposed method is still able to increase the performance. This increase is due to the fact that changes in the x-axis bias are more significant than those due to the rate of change because changes in the x-axis bias normally result in larger errors. Also, since the model to estimate the x-axis bias (Figure 3.7(a)) fits better the parameter, the model is able to compensate for those large errors increasing the performance of the metric. However, this increase is not significant compared with the performance of the other tested VQMs.

Figure 3.9(d), (e) and (f) shows the scatter plot of RMSE and MAE computed for the considered test sequences where the value of 0 means no difference between the tested quality measures and the DMOS. Here, the closer the data points to the bottom left corner, the better the VQM performance. For instance, the best performing methods according to the plots are SOVQM followed by VQAD and CPSNR. Comparing the MAE of the best performing measure using the proposed method (CPSNR) and the other considered quality measures, we found that the proposed method is competitive with SOVQM as well as VQAD. The MAE between DMOS and the pDMOS obtained using CPSNR, SOVQM as well as for VQAD is lower than 0.1. That is, the predicted

DMOS computed with one of those quality measures is expected to be deviated ±10% from its real value. Therefore, the results show that we can perform as well as the state-of-the-art methods to predict quality of corrupted sequences *by using a very simple measure*.

As expected, the performance of the SOVQM and VQAD metrics is higher than the performance of the CSOVQM and CVQAD because the estimated model does not accurately express the relationship between VQM and DMOS under different source sequence (cf. confidence intervals Figure 3.5 and 3.8). The decrease in performance of CSOVQM and CVQAD with respect to their non-content-aware counterparts is mainly due to the poor generalization power of the selected models for these measures. For instance, by exploring the PCC in the training phase, we have PCC = 0.85 for the SOVQM and PCC = 0.86 for the CSOVQM. That is, there is an increase in performance by using content information in the training samples. However, in the testing phase we have PCC = 0.85 for the SOVQM and PCC = 0.82 for the CSOVQM. This shows that there is an increase in performance by adding degrees of freedom (from two parameters for SOVQM to six for CSOVQM) to the fitting function (training results) but it also suggests a poor generalization power of the model due to the mechanisms to mimic the visual system used by these VQMs (testing results).

The results of the proposed methodology on SOVQM and VQAD show that it is difficult to find a model for predicting the parameters of the mapping function to compensate for content information. This is mainly because those metrics use complex mechanisms to mimic the visual system increasing the complexity of the relationship VQM→DMOS under different source sequence, thereby increasing also the complexity of the modeling procedure. This can be a disadvantage because there is no guarantee of finding a model for every VQM. Therefore, the study of different strategies to address this content dependency is proposed as future work. For example, the parameter estimation can be improved by using higher order models or different model estimation methodologies (e.g., support vector regression). Additionally, it would be of interest to minimize the errors between p(D)MOS and (D)MOS during the parameter estimation instead of solving the problem of Equation (3.2).

In summary, for PSNR is easy to model its relationship with DMOS under different source sequence but it is more difficult to model such a relationship for the other VQMs. The results suggest that TPB methods are more suitable for the proposed method than for the other tested VQMs. Since the `IRCCyN` and `IVP` databases were used during the *content related indices* selection, we use other three sets of data to validate our method with the purpose of avoiding cross-validation errors. In particular, we use the `IRCCyN IVC Influence Content` database (Pitrey et al., 2012) as well as the `CIF` and `4CIF EPFL-PoliMI Video Quality Assessment Database` (De-Simone et al., 2009). Note that, these databases were not used in the entire process of model training, *content related indices* and/or mapping functions selection, i.e., we use a training, validation and test sets to measure the performance of the proposed method.

Table 3.2: Performance of considered video quality metrics appraised on IRCCyN IVC Influence Content as well as CIF and 4CIF EPFL-PoliMI Video Quality Assessment Database.

IRCCyN IVC Influence Content				CIF EPFL-PoliMI				4CIF EPFL-PoliMI			
Method	**RMSE**	**MAE**	**PCC**	**SROCC**	**RMSE**	**MAE**	**PCC**	**SROCC**	**RMSE**	**MAE**	**PCC**	**SROCC**
PSNR	0.118	0.091	0.833	0.837	0.254	0.208	0.601	0.692	0.218	0.176	0.676	0.764
SSIM	0.190	0.161	0.462	0.508	0.325	0.279	0.697	0.722	0.293	0.255	0.731	0.750
SOVQM	0.082	0.062	0.941	0.916	0.139	0.115	0.933	0.927	0.230	0.203	0.895	0.928
VQAD	0.073	0.056	0.924	0.942	0.149	0.125	0.913	0.918	0.110	0.094	0.951	0.967
CPSNR	0.095	0.072	0.896	0.891	0.162	0.123	0.865	0.854	0.150	0.121	0.879	0.892

Since we have shown that the proposed method has major advantages in TPB methods, we further validate only the proposed method in PSNR and we compare with the other tested VQMs. That is, we compare between the following methods PSNR, SSIM, SOVQM, VQAD and CPSNR on the IRCCyN IVC Influence Content database (Pitrey et al., 2012) as well as on the CIF and 4CIF EPFL-PoliMI Video Quality Assessment Database (De-Simone et al., 2009). The results of this test are shown in Table 3.2. The performance appraised on the IRCCyN IVC Influence Content database is high for most of the tested VQMs except the SSIM. This can be due to the fact that SSIM is computed frame by frame and the global quality measure is given by the average over all frames which can lead to big estimation errors because it is well-known that the average is highly affected by the distribution of the data which may not take into account the PVQ distribution across time (Keimel et al., 2010). We attribute the good performance of the other quality measures (even PSNR with PCC equal to 0.833) to the fact that the motion distribution of the sequences is not very diverse. In fact, most of the sequences are located within a small interval of TA (TA lower than 15, cf. Figure A.10) compared with the other databases. This makes computing the predicted quality measures easier because the more similar the *content related indices* between the sequences the more similar the parameters of the mapping function. That is, only one mapping function would be necessary to fit the data points. However, by using the proposed method (the same model as estimated using IRCCyN and IVP databases, cf. Section 3.3.2), we can still achieve higher performance than PSNR with an increase of 7.5% (21% in linear Fisher's Z) in PCC as well as in

Table 3.3: Time ratio with PSNR. The time ratio is computed for algorithms running in Matlab (Laptop with CPU intel core i3 2.27GHz and 4GB ram) for 250 frames of size 768×432.

Method	Time ratio	Method	Time ratio
PSNR	1	wTQM	120
SOVQM	16	VQAD	8
wSSIM	246	CPSNR	2

SROCC and a decrease of 20% in RMSE as well as in MAE.

We can draw a similar conclusion by exploring the results on the `CIF` and `4CIF EPFL-PoliMI` databases (cf. Table 3.2). PSNR and SSIM are still the worst performing quality measures because their results are highly variable from database to database, i.e., there is little generalization power on these two VQMs. We also found that CPSNR performs better than PSNR increasing the PCC as well as SROCC in 16% (42% in linear Fisher's Z) and a decrease higher than 30% in RMSE as well as in MAE. These results agree with the results shown in Figure 3.9. Furthermore, since these databases (`IRCCyN IVC Influence Content`, `CIF` and `4CIF EPFL-PoliMI`) contain video sequences with spatial as well as temporal resolutions different from the ones used to train the CPSNR model, and the distortion types are different compared with the `IRCCyN` and `IVP` databases (see the database descriptions in Appendix A.5), we can conclude that the proposed method and the parameters obtained during the training phase work under different scenarios. This indicates that the proposed method can be used over different range of spatial and temporal resolution as well as distortions types. However, the effect of changing the type of distortion requires further study because in the `CIF` and `4CIF EPFL-PoliMI` databases only a simulation of packet loss over the sequences compressed using H.264 codec is added compared with the compressed sequences using the same codec in the training data. That is, there is a strong relationship between the two types of distortion in these databases. Therefore, more research is necessary to determine if the proposed method is able to handle different distortions types.

Table 3.3 shows the computational time measured in Matlab using a Laptop with CPU intel core i3 2.27GHz and 4GB ram for 250 frames of size 768×432. Even though the proposed method does not have the highest performance among the tested methods, the CPSNR computational time is only 2 times higher than the PSNR (being the PSNR the fastest method [see Table 3.3]). This is a low computational time compared with the other state-of-the-art VQMs which use 8 times (or more) the computational time of the PSNR. That is, the computational time used for the *content related indices* is comparable with the computational time used for PSNR. We further study the computational time of CPSNR in Appendix C. Thus, the proposed method (CPSNR) has lower computational complexity compared to SOVQM, VQAD and other

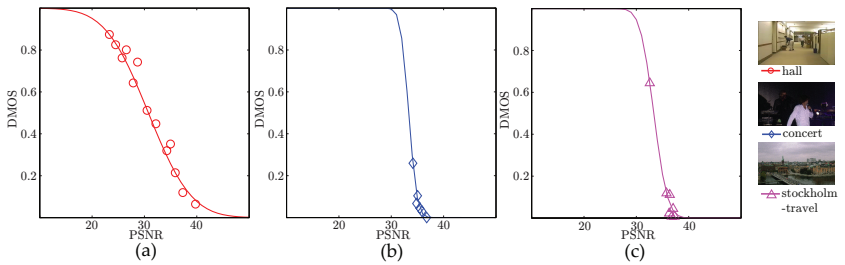

Figure 3.10: DMOS in function of PSNR. Each marker symbol represents a different reference source sequence (a) hall (De-Simone et al., 2009), (b) concert (Pechard et al., 2011), (c) stockholm-travel (Pechard et al., 2011). The solid lines represent curves with the best fit to the data. (a) Example of well distributed subjective quality scores, (b) and (c) examples of subjective quality scores distributed over small perceived quality region.

more sophisticated methods such as MOVIE, wTQM, wSSIM (Li et al., 2011b; Ortiz-Jaramillo et al., 2014a). This is a major computational advantage because the method is based on very simple operations which are used to characterize the content of the video sequence instead of computing more complex features in local blocks (SSIM and SOVQM) or trying to mimic the human visual system (MOVIE, wTQM and wSSIM) which in general is computationally more complex (Li et al., 2011b; Ortiz-Jaramillo et al., 2014a). The results show that the proposed method (CPSNR) is competitive with current state-of-the-art VQMs as far as prediction accuracy is concerned. However, this good performance is achieved with a significantly reduced computational time.

The main drawback of the proposed method is that an off-line training with enough samples representing the wide range of quality levels, extent of details and motion is needed. That is, the performance of the method depends highly on the variety of video content included during the training phase. This issue is currently difficult to address due to the lack of public databases fulfilling such requirements (Winkler, 2010).

3.4.2 Selecting test sequences for subjective experiments

When designing a subjective study for video quality assessment, preparation of corrupted video sequences (test stimuli) to be rated by human subjects is a challenging task because they affect the usefulness of the collected human data. This usefulness is reflected by whether or not the resulting (D)MOS scores are uniformly distributed over its entire range, which depends completely on the selected acquisition, processing and technical parameters (Kumcu et al., 2015b). It is known that such parameters (e.g., noise, blur, compression rate, PSNR value, among others) are often non-linearly related to PVQ and the model of the relationship may be unknown a-priori.

Figure 3.10 shows plots of DMOS in function of PSNR for different cases of

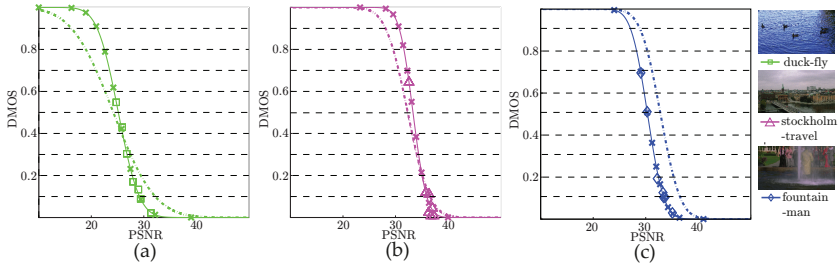

Figure 3.11: DMOS in function of PSNR. Each marker symbol represents a different reference source sequence (a) duck-fly, (b) stockholm-travel, (c) fountain-man sequences (Pechard et al., 2011). The solid lines represent curves with the best fit to the data. The dotted lines represent curves estimated by using the proposed method. The crosses are projections of PSNR values selected by using the estimated curve on to the mapping function specifically tuned to the source sequence. Horizontal dashed lines divide the DMOS axis equally in steps of 0.1.

source sequences, taken from `IRCCyN` and `CIF EPFL-PoliMI` databases. These examples illustrate the drawbacks of current selection of distortion levels for subjective studies, which are mainly due to the lack of standard procedures for this selection. For instance, Figure 3.10(a) shows an example of subjective quality scores distributed over the (quasi)linear range of the relationship DMOS-PSNR. Note that this is not yet an optimal selection of test sequences because there are corrupted sequences with almost the same DMOS for the same source sequence and it would be more desirable to have DMOS values in the saturation range as well (Winkler, 2010; Kumcu et al., 2015b) (e.g. DMOS$>$ 0.9). In Figure 3.10(b) and (c) quality levels are almost exclusively located in the low saturation range, i.e., the data has too little variety in DMOS.

To address the problem of adequate parameter selection for the test stimuli, (Kumcu et al., 2015b) have proposed a method for modeling the relationship between parameter levels and PVQ using a paired comparison procedure in which subjects judge the perceived similarity in quality (Kumcu et al., 2015b). Their results indicate that the obtained subjective scores were well distributed over the entire DMOS range. Nevertheless, that method requires a small subjective pre-study (pilot study) for modelling the relationship between parameter levels and PVQ. This can be a disadvantage because, although it is a small experiment, it is still time consuming and highly subjective for the initial selection of the distortion levels. Instead, we propose to use CPSNR for the selection of the distortion levels because a pilot study is not necessary. In the following paragraphs we give a detailed description of our proposed method of parameter selection for the test stimuli by using some test samples and PSNR as technical measure.

For a given source sequence, it is possible to select a uniformly sampled DMOS domain by applying the following steps: (i) compute $\mathbf{s}$ and $\mathbf{t}$, (ii) com-

pute $\hat{\mathbf{a}} = A[1, \mathbf{s}, \mathbf{t}]^T$, (iii) divide the DMOS axis equally using the desired step size, and (iv) use the estimated parameters $\hat{\mathbf{a}}$ as well as the divided DMOS axis to obtain the corresponding PSNR values, i.e., the appropriate set of distortion levels. Thereafter, distorted sequences are generated to correspond to these PSNR values.

To illustrate the method, we use three source video sequences taken from `IRCCyN` database together with their distorted versions. Figure 3.11 shows examples of DMOS in function of PSNR for the different reference source sequence (a) duck-fly, (b) stockholm-travel, (c) fountain-man sequences. The markers (squares, triangles and diamonds) are the scatter plots of DMOS scores from the `IRCCyN` database and the corresponding PSNR values computed between the source (reference) and the distorted sequences. The solid lines represent curves with the best fit to the data points, i.e., the "true" model between PSNR→DMOS. We call it "true" model because the mapping function was specifically tuned to the source content using the DMOS obtained through subjective evaluation. The dotted lines represent curves estimated by using the proposed method, i.e., we use the model shown in Figure 3.6 and the *content related indices* extracted from the example sequences to compute the parameters of the mapping function PSNR→DMOS. To obtain a roughly equally sampled DMOS space, we divide the DMOS axis equally in steps of 0.1 (see horizontal dashed lines in Figure 3.11). Then, we use the curves represented by the dotted lines to obtain the preferred PSNR values that should be obtained between a distorted sequence and the given reference sequence.

Figure 3.11(a) and (b) show two examples where the proposed method was able to recommend PSNR values that divide the DMOS domain equally as it is desirable (Winkler, 2010; Kumcu et al., 2015b). Figure 3.11(c) shows an example in which the proposed method does not divide the DMOS domain equally, i.e., the points are not equally distributed over the whole perceived quality range. In any case, the plots show that the selected values using the proposed method (crosses) are more uniformly distributed in the DMOS axis than the ones selected in the original database (squares, triangles and diamonds).

To illustrate the potential of this method, we use the experimental CPSNR data shown in Table 3.2 MAE columns, i.e., the MAE value achieved by using the CPSNR tested on the `IRCCyN IVC Influence Content` as well as the `CIF` and `4CIF EPFL-PoliMI` databases. Even though the MAE is estimated by using the model trained with `IRCCyN` and `IVP` databases, the CPSNR has MAE value of 0.072 when the distortion type is the same as in the training phase (`IRCCyN IVC Influence Content` [compression with H.264 codec]) and 0.12 when the distortion type is different (`CIF` and `4CIF EPFL-PoliMI` [compression with H.264 codec + packet loss]). This suggests that the expected error between the obtained DMOS using the recommended PSNR and the DMOS that is going to be obtained through the subjective evaluation is ±7.2% and ±12% from its real value when the distortion type is the same as the training and the distortion type is different from the training, respectively.

3.5 Conclusions

In this Chapter we proposed a method to advance existing VQMs by introducing *content related indices* in their computation. The proposed method is based on observations made from the changes of VQMs in function of (D)MOS under varying content. In this work PSNR, SSIM, SOVQM, as well as VQAD and statistics of images filtered with SI13 filter, temporal gradients as well as spatial dependencies of pixel values were used as VQMs and *content related indices*, respectively, with the purpose of illustrating the potential of the proposed method. In particular, our method involves the off-line training of the parameters of the complementary error function. We have found that the linear combinations of spatial and temporal activity (SA, TA) are good predictors of such parameters. However, we have also found that when the VQM includes some mechanisms to mimic the human visual system, it is more difficult to model the changes in predicted quality.

The results show that our method performs well over multiple types of video content, spatial and temporal resolutions. Experiments over five different public video quality databases demonstrate that the proposed method is competitive with current state-of-the-art methods. Also, since the proposed method is based on simple operations, it has shown to be faster and simpler than current state-of-the-art methods. Moreover, since many video-related systems today already rely on PSNR to perform video quality estimation, the proposed PSNR-based method is easy to incorporate into those systems. Additionally, the proposed method has been shown to be generic for including different nonlinear functions, video quality measures, and/or video *content related indices*. Also, CPSNR has shown to be of particular interest because it is possible to estimate PSNR→DMOS curves that can be used to preselect the levels of video distortion in the preparation of subjective studies.

Another contribution of this chapter is the evaluation of four of the most well-known state-of-the-art VQMs (PSNR, SSIM, SOVQM and VQAD) on five different public video quality databases. This is a major contribution because even though these VQMs are the most well-known and widely used, the VQMs are often tested on databases with only a small variety in content, few testing samples and/or data which is not publicly available. Here, we have included databases with notably more testing samples than other work currently presented in the state-of-the-art: 696 distorted sequences from 102 source sequences under different temporal and spatial resolutions. That considered, the results presented in this work can be used as a reference when evaluating newly developed VQMs.

The contributions reported in this Chapter resulted in two international conference proceedings (Ortiz-Jaramillo et al., 2014a, 2015c), and one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2016b). Additionally, this work has been successfully demonstrated in the Imec Technology Forum 2017 (cf. Appendix C) attracting interest from several companies and in relation to several use cases; the follow-up discussions are ongoing.

4

Evaluation of contrast ratio changes in images

4.1 Introduction

In this Chapter image differences are accounted in terms of perceived contrast ratio changes between the test image and the reference image (standard sample). Contrast is a concept that has an intuitive meaning but that has been defined in different ways depending of the context. Particularly, in the image fidelity assessment field contrast is defined as the visual property that makes a structure of interest distinguishable from the background. This is a very important characteristic of many image based systems. For instance, during interventional X-ray image acquisition, specific areas (regions of interest) of the acquired/current image are analyzed to determine the detectability/visibility of the diagnostically relevant details (foreground). Afterwards, the perceived (subjective) visual fidelity of the current image is estimated by comparing the detectability values of the current image to those of the standard (reference) image samples where, according to the interventionalists, the diagnostically relevant details are presented under "ideal" detectability conditions. Here, the contrast ratio is used to compute the visibility of the foreground given the background. Another example is in multiview imaging where color correction (Fezza et al., 2014) and contrast enhancement techniques (Palma-Amestoy et al., 2009; Bertalmio et al., 2009) are often used to adjust the contrast, brightness and/or color settings of the cameras and/or displays. It is well-known that the behavior of these techniques highly depends on the accuracy of the computation of image features such as contrast ratio (Palma-Amestoy et al., 2009). Therefore, it is crucial to have an accurate measure of contrast ratio for improving and evaluating color correction and contrast enhancement techniques. In this case, contrast ratio values are used to compare selected areas of a test image with respect to the contrast values of certain defined reference image, e.g., an image acquired under "ideal" conditions. For example, the contrast ratio values of every view in a multi-camera system are typically

compared and adjusted with respect to the contrast ratio values of the image acquired with the camera defined as the reference camera (Zhao et al., 2013).

There is no standard procedure to measure contrast ratio in images (Panetta et al., 2013). Hence, several measures have been proposed (Panetta et al., 2013; Pedersen et al., 2008). For example, in (Agaian et al., 2000), (Agaian et al., 2007) and (Panetta et al., 2013) measures based on conventional contrast ratio formulas were proposed. Those measures use Simple, Weber's and Michelson's contrast ratio formulas on non-overlapping image patches to compute the contrast ratio index for a given image, termed Simple contrast based Measure of Enhancement (SME) (Panetta et al., 2013), Weber contrast based Measure of Enhancement (WME) (Agaian et al., 2000) and Michelson contrast based Measure of Enhancement (MME) (Agaian et al., 2007). Another methodology is to decompose the image into a pyramidal structure by filtering. Thereafter, the contrast ratio is estimated by dividing each pixel value of the filtered images by the average intensity of the current decomposition level (Peli, 1997), termed Peli's contrast. A very popular and efficient implementation of Peli's contrast is the Wavelet implementation presented in (Provenzi and Caselles, 2014), here termed Peli's wavelet contrast measure (PWC).

In general, the state-of-the-art measures are based on measuring the relative difference between dark and light intensity points of local image patches and/or image sub-bands (Pedersen et al., 2008). However, such techniques fail to accurately compute the contrast ratio when complex backgrounds are present (e.g. highly textured images) because it is known that the detectability/visibility of a human observer is influenced by the surrounding local content (Pedersen et al., 2008; Panetta et al., 2013). That is, the contrast ratio is influenced by the local distribution of pixel values (Provenzi and Caselles, 2014). In general, the local content of an image patch is classified either as flat, textured or edge. Flat areas are of not interesting because there are no changes in intensity. Also, human observers detect more easy contrast changes in patches with local intensity discontinuities (edges) than in other image patches, e.g., textured or flat patches (Provenzi and Caselles, 2014).

Therefore, we propose to estimate the contrast ratio in local image patches by taking into account the local changes around the edges. We use Weber and Michelson contrast ratio formulas on each patch to simulate the cases where a small structure of interest (edge) is present on a uniform background or a square-wave grating of one cycle, respectively (Zuffi et al., 2007). Although in practice the background is typically not uniform, we have found that the edges in local image patches can be characterized by bimodal histograms representing a set of pixels likely to be inside the foreground (edge pixels) and another set likely to be in the background. Then, the local contrast ratio can be estimated using the ratio between mean intensity values of each mode of the histogram. This process is performed over the entire image with a sliding window resulting in a contrast ratio per pixel, termed *contrast ratio map*. Thereafter, statistics of the *contrast ratio map* are used as a overall contrast ratio index.

We have tested our measure on two image quality assessment databases

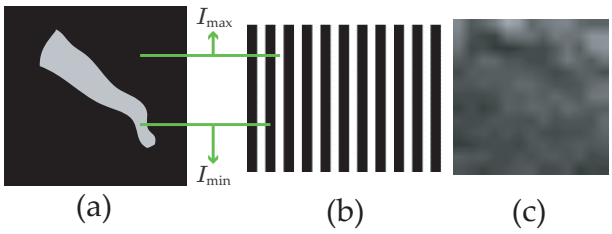

Figure 4.1: Images used in standard definitions of contrast ratio. (a) Weber's and Simple, (b) Michelson's and (c) root mean squared contrast ratio formulas.

(TID2013 (Ponomarenko et al., 2015) and CSIQ (Larson and Chandler, 2010)) to demonstrate that the proposed measure is able to accurately predict image changes due to contrast decrements and increments typically reported by a human observer. Our experimental results show that the proposed method agrees with human judgment (correlation between the subjective scores and the proposed measure exceeds 90%). Additionally, we have tested our methodology on a real case scenario (detection of changes in contrast level in interventional x-ray images acquired at varying radiation dose) (Ortiz-Jaramillo et al., 2015b,a). Particularly, we used two static anthropomorphic chest phantoms scanned at six dose levels simulating a small and a large chest. The results show that the proposed contrast ratio measure agrees well with the subjective differences reported by the interventionalists (cardiologist/radiologist).

The rest of the Chapter is organized as follows. Section 4.2 introduces background information and Section 4.3 describes the proposed methodology. The experimental setup and results are presented in Section 4.4. Finally, in Section 4.5 conclusions are outlined.

4.2 Background

The visual property that makes a structure of interest distinguishable from the background is typically quantified by the contrast ratio, i.e., the visibility of the foreground given its surrounding background. Therefore, many formulas of contrast ratio have been proposed, e.g., Simple, Weber's, Michelson's, root-mean-squared contrast ratio formulas, among others (Zuffi et al., 2007; Pedersen et al., 2008; Panetta et al., 2013).

4.2.1 Classic definitions of contrast

In the following paragraphs we describe the classic definitions of contrast ratio.

Weber's contrast ratio formula is applied in cases where a small structure of interest is present in a uniform background such as in Figure 4.1(a) and it

is computed as

$$c_{\mathrm{W}} = \frac{I_{\max} - I_{\min}}{I_{\max}}.$$

Simple contrast ratio formula is a simplification of Weber's contrast ratio formula that is used in cases where the viewer is assumed to have adapted to the background intensity, i.e.,

$$c_{\mathrm{S}} = \frac{I_{\min}}{I_{\max}}.$$

Michelson's contrast ratio formula is commonly used for patterns where both bright and dark take up similar fractions of the area under inspection (e.g. the square-wave grating in Figure 4.1(b)) and it is computed as

$$c_{\mathrm{M}} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}},$$

where $I_{\max}$ and $I_{\min}$ denote the maximum and minimum intensity of the area under inspection, respectively (Zuffi et al., 2007; Panetta et al., 2013). The root-mean-squared contrast (RMSC) is computed using the mean squared error between the central pixel and K surrounding neighbors as defined by (Rizzi et al., 2004)

$$\mathrm{RMSC}_1 = \sqrt{\frac{1}{K}\sum_k \left(I_{\mathrm{center}} - I_k\right)^2}.$$

Another RMSC measure uses the ratio between the difference and addition of the average value and the central pixel as (Panetta et al., 2013)

$$\mathrm{RMSC}_2 = \frac{|I_{\mathrm{center}} - \frac{1}{K}\sum_k I_k|}{|I_{\mathrm{center}} + \frac{1}{K}\sum_k I_k|}.$$

In either case, the RMSC is a measure of variability of the pixel values with respect to the central pixel proposed for random patterns like in Figure 4.1(c). Therefore, RMSC is not a true contrast ratio measure but rather it is an approximation based on the deviation of the pixel values.

4.2.2 Contrast ratio measures in images

Traditionally, contrast ratio in images is determined based on measuring the contrast ratio of local image patches by using one of the previously defined formulas. Thereafter, the obtained values are averaged to obtain an overall contrast ratio index. Some of the most popular contrast ratio measures used in images in the state-of-the-art are listed in Table 4.1. In the following paragraphs we describe the computation of these measures.

SME, WME, MME and RMS compute, respectively, Single, Weber's, Michelson's and RMSC_2 contrast ratio formulas on non-overlapping blocks. These measures also compute the logarithm of each estimated contrast ratio value to simulate the logarithmic sensation of the human eye (Panetta et al.,

Table 4.1: List of the tested contrast ratio measures in images

Name	Acronym
Simple contrast measure of enhancement	SME (Panetta et al., 2013)
Weber's contrast measure of enhancement	WME (Agaian et al., 2000)
Michelson's contrast measure of enhancement	MME (Agaian et al., 2007)
Root mean squared measure of enhancement	RMS (Panetta et al., 2013)
Multi-scale RMS	MSRMS (Rizzi et al., 2004)
Peli's wavelet contrast measure	PWC (Provenzi and Caselles, 2014)

2013). MSRMS applies the same methodology as RMS but uses a multi-scale approach and the RMSC_1 formula. That is, the image is decomposed in a pyramid by subsampling. The image is decomposed in three levels and thereafter the RMSC_1 formula is computed block-wise independently on each band. PWC decomposes the image into a pyramidal structure by filtering (Peli, 1997). First the image is decomposed using a set of filters. A very efficient way to do so is by using the discrete wavelet transform (DWT) (Provenzi and Caselles, 2014). Afterwards, the approximation coefficients from each decomposition level of the DWT are divided by their corresponding detail coefficients obtaining a local contrast measure at each level. The global contrast ratio index in all these methods is computed as the average of the resulting local values.

The traditional contrast ratio measures either assume background uniformity (SME, MME and WME) or compute contrast ratio using the local pixel values deviations (RMS and MSRMS). In general, these techniques fail in computing the contrast ratio under complex backgrounds. On the one hand, the background uniformity often cannot be assumed because natural scene images possess a wide range of different texture features and the inherent noise in images used in medical applications, for example interventional x-ray (Kumcu et al., 2015a,b; Ortiz-Jaramillo et al., 2015b,a), cannot be avoided. On the other hand, the pixel value deviations often provide more information about texture or noise level than contrast ratio. Note that none of the contrast ratio measures take into account the fact that human observers perceive easily contrast ratio changes around the local intensity discontinuities or edges (Agaian et al., 2000; Panetta et al., 2013; Provenzi and Caselles, 2014). Therefore, we propose to compute contrast ratio around the edges while considering the local distribution of pixel values.

4.3 Proposed method

In this section, we first study image content characteristics for describing the local distribution of pixel values, i.e., the distribution of pixel values around image edges. Afterwards, we provide our local contrast ratio definition based on these characteristics.

In general, image content is studied from two perspectives: globally (Pinson and Wolf, 2004a) or locally (Thung et al., 2012). On the one hand, the global perspective computes features globally to represent the whole image content,

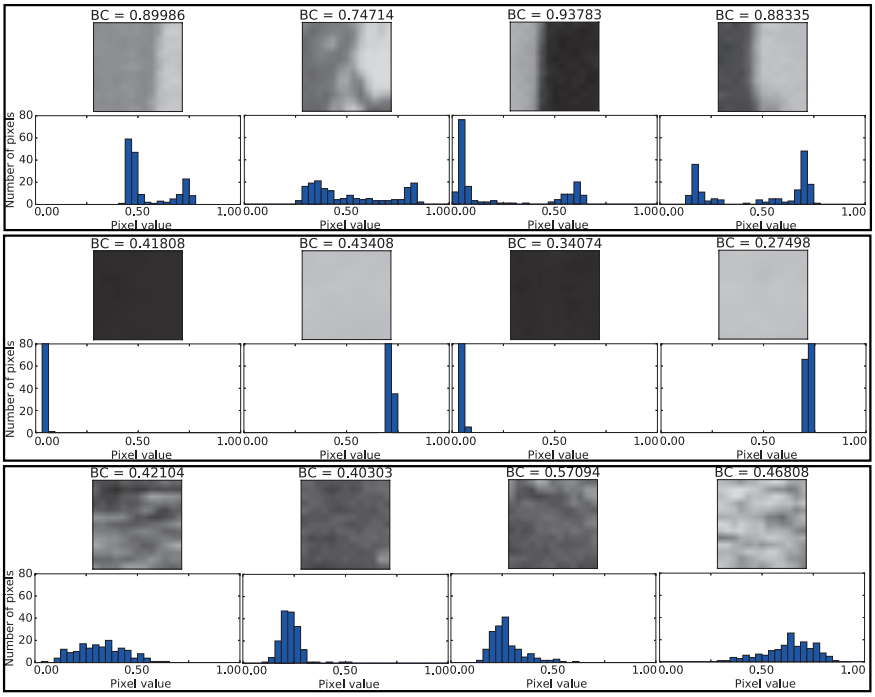

Figure 4.2: Random patches of 13×13 pixels from the "miscellaneous" set of the USC-SIPI Image Database and their corresponding histogram. From top to bottom edges, flat and textured patches.

e.g., descriptive statistics of the magnitude of the image gradient. On the other hand, the local perspective classifies a small image region or patch either as flat, textured or edge (Thung et al., 2012). Figure 4.2 shows examples of flat, textured and edge patches. Note that the local perspective of content analysis will be used for computing contrast ratio around the edges. We recommend the work of (Pinson and Wolf, 2004a) as well as Chapter 3 to the readers interested further in the global perspective.

4.3.1 Local content analysis

We have investigated multiple small local patches with the purpose of characterizing the local distribution of pixel values using histograms. Particularly, we have extracted 250 random patches of 13×13 pixels from the "miscellaneous" set from the USC-SIPI Image Database (airplane, baboon, boat, bridge, cameraman, house, lake, lena, man, peppers [25 per image]) (USC Viterbi, 2016). We use a 13×13 pixels window because significant edges are accurately detected by using a 13 pixels high pass filter (Pinson and Wolf, 2004a) meaning that this size is an appropriated selection for studying edges. The 250 image

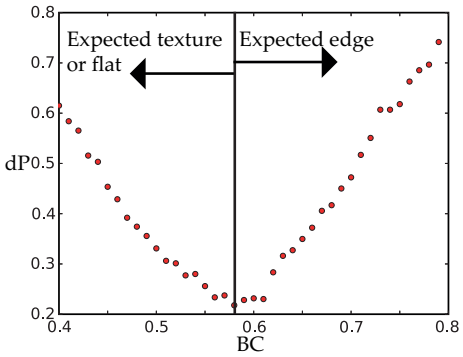

Figure 4.3: Scatter plot of the BC thresholds and the distance to the perfect classifier (dP).

patches were manually labeled by a human observer as edge (structure of interest) or textured/flat. Figure 4.2 shows examples of the used patches. Here, the bimodality coefficient (BC) is computed to assist in the modeling of the histograms (Pfister et al., 2013). The BC is a measure to distinguish between unimodality and bimodality. The BC is based on an empirical relationship between bimodality and the third and fourth statistical moments of a distribution with the underlying logic that a bimodal distribution will have either a very low fourth statistical moment, or a very high third statistical moment, or both (all of these conditions increase BC). The BC is defined as follows (Pfister et al., 2013):

$$\mathrm{BC} = \frac{\mathrm{skew}^2 + 1}{\mathrm{kurt} + \frac{3(n-1)^2}{(n-2)(n-3)}},$$

where skew, kurt and n are the skewness, kurtosis and number of pixels in the image patch, respectively. The BC value ranges between 0 and 1. We use the BC to automatically classify each patch as edge or textured/flat by thresholding it in the range of 0.4 - 0.8. This specific range is selected empirically because the values outside this interval lead to poor classification rates as the trend shows in Figure 4.3. Thereafter, we compute the distance to the perfect classifier (dP) for each threshold as:

$$\mathrm{dP} = \sqrt{(1 - \mathrm{TPR})^2 + \mathrm{FPR}^2},$$

where TPR and FPR are true positive and false positive rates, respectively.

Figure 4.3 shows the scatter plot of the thresholds (BC values) and the dP. The plot shows that the highest performance (lowest dP) is achieved using a threshold BC belonging to the interval $[0.55, 0.60]$ with an accuracy ranging between 80 and 90%. The obtained BC interval is of particular interest because the threshold value to expect a uniform distribution is BC= $\frac{5}{9} \approx 0.555$. In practice, higher values of BC point toward multimodality whereas lower values

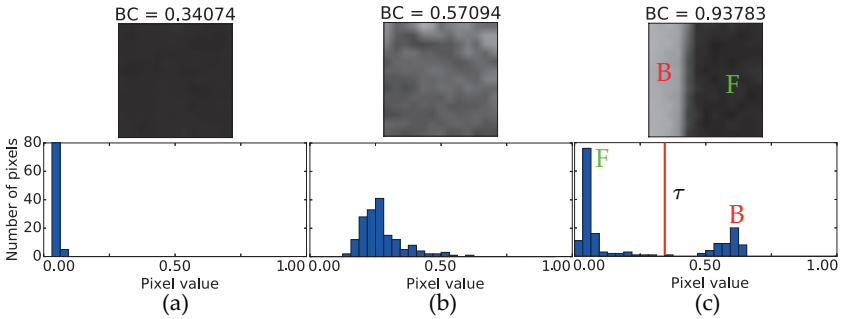

Figure 4.4: Types of local content in small image patches (top row) and their histograms (bottom row). (a) Flat, (b) textured and (c) edge patch. F, B and τ represent the foreground pixel intensities, the background pixels intensities and the mid-point between the two modes of a histogram, respectively.

point toward unimodality (Pfister et al., 2013). This analysis shows that the local histogram of a structure of interest (edge) is, unlike textured and flat areas, multimodal distributed. In general, the experiments reveal that while textured/flat patches can be characterized by unimodal histograms (see Figures 4.2 (middle-bottom) and 4.4(a)-(b)), edges are characterized by bimodal histograms representing the background and foreground (see Figures 4.2 (top) and 4.4(c)).

4.3.2 Content-aware contrast ratio

We propose to estimate the contrast ratio by using Weber's and Michelson's formulas in local image patches taking into account the bimodal property of the histograms of edges as explained in Section 4.3.1. We have used Weber's and Michelson's contrast ratio formulas because they can be used in cases where a small structure of interest is present on a uniform background and a square-wave grating of one cycle, respectively (Zuffi et al., 2007). Although in practice the background is non-uniform, we have shown that an edge is characterized by bimodal histograms where each mode of the histogram represents a distribution likely to be in the edge (foreground) and another distribution likely to be in the background. Then, the local contrast ratio is estimated using the ratio between average intensity values of each distribution, i.e.,

$$\begin{aligned} c = & \quad 1 - \frac{f}{b} \quad \text{and} \\ c = & \quad \frac{|b-f|}{b+f} \end{aligned} \tag{4.1}$$

for Weber's and Michelson's formulas, respectively. Here c, f and b are the contrast ratio, the average foreground intensity and the average background intensity, respectively. f and b are computed after identifying the edge (fore-

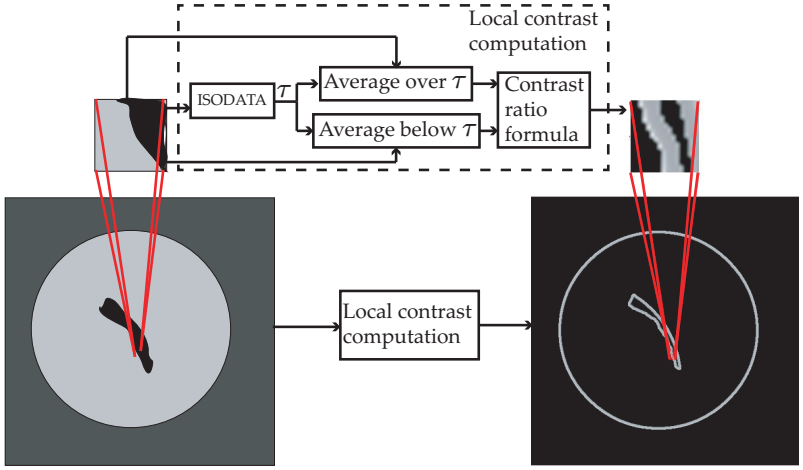

Figure 4.5: Block diagram of the proposed contrast ratio measure.

ground) and background in the image patch by computing the mid-value between the modes of the histogram using the `ISODATA` algorithm (Dianat and Kasaei, 2008). The `ISODATA` algorithm is described in Appendix D.

After obtaining the threshold τ, the corresponding background and foreground intensities (f and b) are computed as the average of each set of pixels. Finally, the contrast ratio is computed by using one of the formulas in Equation (4.1).

Figure 4.5 shows the block diagram of the proposed local contrast ratio measure for images. The computation is performed block-wise over the entire image using a sliding window (Local contrast computation block). For each image patch visited by the sliding window, the `ISODATA` algorithm returns a threshold (τ) used next to compute the average values of the foreground and background (average below- and over the threshold) to later estimate the local contrast ratio by using one of the formulas in Equation (4.1). In this thesis we test independently Michelson's and Weber's formulas into our methodology. The result is a contrast ratio value per pixel or per block (for non-overlapping sliding windows), termed contrast ratio map. Finally, the local contrast values are aggregated into a single global contrast estimate for the whole image by percentile pooling, as detailed in Section 4.3.3.

Note that the contrast ratio values are computed over the entire image without discriminating if the patch is a structure of interest (edge) or just background (textured or flat) with the purpose of keeping a low computational time. That is, we do not find the structure of interest deterministically but locate a set of pixels which are likely to be inside the foreground and another set likely to be in the background. This can lead to potential estimation errors mainly due to the used contrast ratio formula. Future work should explore potential

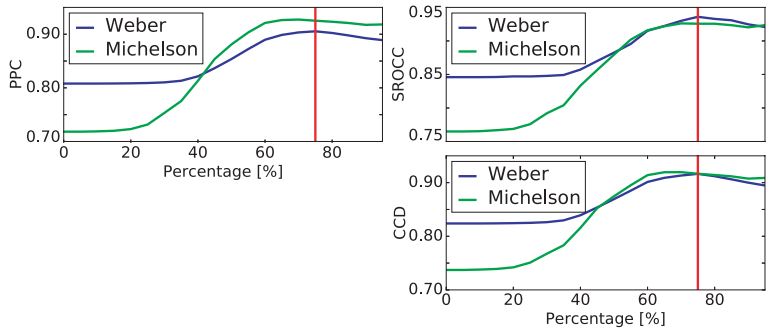

Figure 4.6: Performance of the considered percentile pooling using different threshold levels on the *contrast ratio maps* appraised on the CSIQ database.

benefits of computing with a different formula the local contrast ratio values in the background patches. For instance, it is more accurate to use RMSC_2 for random patterns like in Figure 4.1(c). Also, it is more accurate to use the Michelson's formula for square-wave gratings (edges) like in Figure 4.1(b). One way to handle this is by computing contrast ratio using Michelson's formula in patches with bimodal distribution and using RMSC_2 in those local patches where no bimodal distribution is found.

4.3.3 Implementation details

After computing the contrast ratio in local patches, it is necessary to estimate the global measure of contrast for the whole image. For this purpose, we use the harmonic mean as a global measure of image contrast ratio because it has been identified as an appropriate measure when the average of ratios is desired (Zar, 2009). Since the human visual system is more affected by changes in the extreme values of contrast, we will consider only the extremes of the computed local contrast ratios (percentile pooling) (Moorthy and Bovik, 2009a,b). The nth percentile of an ordered set is the highest $100-n\%$ values of that set. Thus, the percentile pooling used in our method first sorts the values in the set in ascending order of the magnitude and then takes the harmonic mean of the highest $100-n\%$ of these values to obtain an overall measure for the image. Particularly, we use the 75% threshold level of the highest values because it has shown very good results in psychophysics (Kingdom and Prins, 2010b). Although this is a theoretically good choice, we explore the impact of percentile pooling of the harmonic mean using different threshold levels in the range 5 to 95% in steps of 5%. Note that the experiments in this Section are performed only on CSIQ database and the same parameters are used for the evaluation on the rest of the tested data. We explore as well the impact of using Weber's and Michelson's formulas.

Figure 4.6 shows the performance (PCC, SROCC and CCD) of the consid-

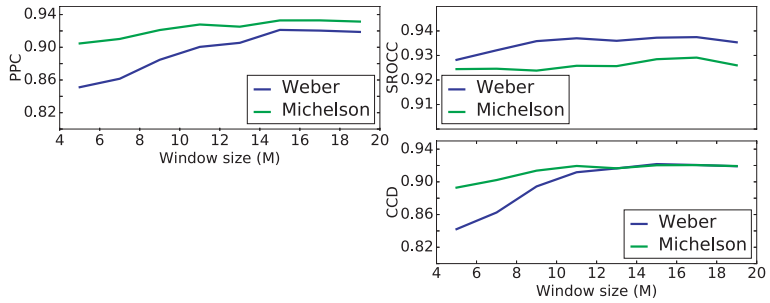

Figure 4.7: Performance of considered patch sizes (M×M) to compute the *contrast ratio maps* appraised on CSIQ database.

ered threshold levels computed on the *contrast ratio maps* using Weber's and Michelson's formulas for the CSIQ database. The graph shows that contrast ratio values lower than the 50% level are irrelevant for the overall contrast ratio difference value, i.e., very low correlation between subjective scores and the numerical contrast ratio differences. Indeed, it is well-known that human subjects are more sensible to changes in regions containing the extreme values (Moorthy and Bovik, 2009b). Also note that the theoretical value (75% threshold level) is the value where the proposed method achieves its highest performance, as expected. Additionally, the results in Figure 4.6 do not show any significant difference in terms of performance between Weber's and Michelson's formulas, at least not around the selected threshold level.

We show as well the impact of the patch size (M×M) in the proposed methodology using Weber's and Michelson's formulas and the 75% threshold level as discussed previously, i.e., we gradually increase the size of the image patches in steps of 2 from 5×5 to 19×19 using the previously selected parameters. When selecting the value of the local patch size parameter of our method, it is important to take into consideration the texture content of the images: the patch size should be chosen large enough to include an edge (i.e. both background and foreground pixels) but not too large to infringe the bi-modality assumption. In general, it is important that the patch size is chosen to approximately satisfy the assumption of bi-modal histograms of the image patches but the method does not require bi-modality of the global image histogram. Accordingly, the proposed method can be used for various types of image content, including natural scene images (as in TID2013 and CSIQ databases) which may have variable and complex global histograms.

Figure 4.7 shows the performance (PCC, SROCC and CCD) of the considered patch sizes (M×M) on the *contrast ratio maps* using Weber's and Michelson's formulas for the CSIQ database. The graph shows that the Weber's formula is more sensitive to changes of the patch size parameter than the Michelson's formula. This is because while the Weber's formula was designed for foregrounds like the one in Figure 4.1(a), which in general needs

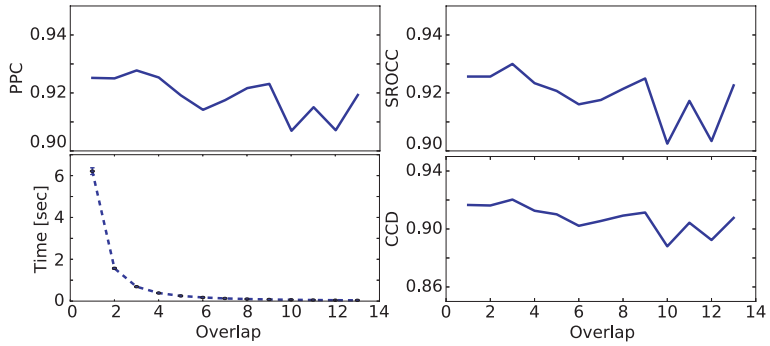

Figure 4.8: Performance of the considered overlap parameters on computing the *contrast ratio maps* using Michelson's formula appraised on the CSIQ database.

bigger window sizes to be obtained, the Michelson's formula was designed for patterns where both bright and dark take up similar fractions like the edges characterized in Section 4.3.1. In any case, the plot shows that M = 15 is a good selection for both formulas because the method achieves its highest performance.

Lastly, we explore the overlap parameter, i.e., the number of overlapped pixels during the sliding window process. We explore overlap parameters ranging from 1 (pixel-wise sliding windows) to 15 (block-wise sliding windows or no overlap).

Figure 4.8 shows the performance (PCC, SROCC and CCD) in function of the considered overlap parameters to compute the *contrast ratio maps* using Michelson's formula for the CSIQ database. The results show that increasing the value of this parameter reduces considerably the computational time. The computational time was measured on functions implemented using Cython (Cython, 2016) on a standard laptop with CPU intel core i7 - 5500U and 12 GB ram running Ubuntu 16.04 LTS. The computational time is the average time in seconds after 20 runs of the algorithm (the average time has a variation of $\pm 5\%$) on an image of 512×384 pixels in size. In the rest of the experiments we use an overlap parameter equal to 3 because it reduces considerably the computational time while maintaining a good performance. Note that each contrast ratio value is computed independently by using a sliding window. Therefore, the result of the computed contrast ratio values depends only on the pixel values of the image patch visited by the sliding window. Thus, the proposed algorithm can be optimized by using parallel computations, i.e. by processing each image patch simultaneously which could lead to big speedups (10 - 50 times compared to single-threaded CPU execution) (Goossens et al., 2014).

Due to the content awareness of the proposed method, we refer to the measures as CWMC and CMMC for content-aware Weber's and Michelson's measure of contrast, respectively.

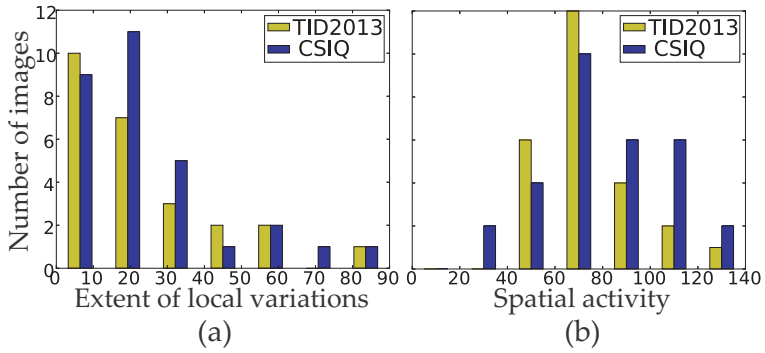

Figure 4.9: Histogram distribution of the (a) extent of local image variation and (b) spatial activity of the reference images in TID2013 and CSIQ databases.

4.4 Results and Discussion

In this Section we describe the used test images, the implementation details of the proposed methodology and the performance comparison with the state-of-the-art measures.

4.4.1 Test images

We test our measure on two image quality assessment databases (TID2013 (Ponomarenko et al., 2015) and CSIQ (Larson and Chandler, 2010)) to demonstrate that the proposed measure agrees with human judgment of perceived contrast changes. We also test our method on interventional x-ray images to evaluate the advantages of the proposed methodology under non-uniform background (textured anatomical and/or noisy background) (Kumcu et al., 2015a,b). These databases are described in more detail in Appendix A. Here we concentrate only on the contrast subset of these databases and their particularities.

Figure 4.9 shows the histogram distribution of (a) the extent of local image variation and (b) the spatial activity computed on the reference images of the considered databases. Each bin of the histogram represents the number of images within a range of extent of local image variation and spatial activity, respectively.

The spatial activity refers to the perceptually significant edges present in the image. For computing the spatial activity of the images we use the average of the magnitude of SI13 (the spatial information SI13 filter is a spatial filter designed specifically to measure the perceptually significant edges by using a 13 pixels highpass filter) filtered images as proposed in (Pinson and Wolf, 2004a):

$$\mathrm{SA} = \frac{1}{NM} \sum_{n,m} |\mathrm{SI13}\{L^*\}(n,m)|,$$

Figure 4.10: Examples of images with contrast decrements: (a) TID2013 (also increments) and (b) CSIQ.

where $|\mathrm{SI13}\{L^*\}(n,m)|$ is the magnitude of the intensity of the image filtered by using the SI13 filter in the (m,n)th pixel. This is a measure of the spatial activity of the perceived details within the image under analysis.

The extent of local image variation is a measure of complexity of the image scene or the amount of texture in the image. The extent of local image variation is computed as the entropy of the co-occurrence matrix defined as

$$\mathrm{ELV} = \sum_{x,y} C(x,y) \log\left(C(x,y)\right),$$

with $C(x,y)$ representing a count of the number of times that $I(n,m) = x$ and $I(n+\Delta n, m+\Delta m) = y$, where $(\Delta n, \Delta m) \in \{(0,1),(-1,1),(-1,0),(-1,-1)\}$, cf. (Randen and Husoy, 1999).

These histograms are an indication of the extent of texture and spatial activity contained in both databases. Overall, the TID2013 subset has fewer images with highly textured areas and/or spatial activity. On the one hand, TID2013 reference images possess little spatial activity suggested by almost half of the SA values located in the lower value range of the histogram. Figure 4.9(b) shows that 18 out of 25 reference images (72%) have SA$<$80, that is low spatial activity, of which 12 (48%) images are located in the narrow SA range of 60-80. This could lead to overestimation of the performance in some of the tested contrast ratio measures as the results will suggest later. On the other hand, the CSIQ images are more spread across the different SA ranges. This is an indication that CSIQ database is more challenging for evaluating contrast compared to the TID2013 database. The two histograms show that TID2013 subset has fewer images with highly textured areas and/or spatial activity.

Figure 4.10 shows examples of the contrast changes available in TID2013 (this database possesses both contrast increments and decrements) and CSIQ, top and bottom rows, respectively. In particular, TID2013 and CSIQ each have a subset of images altered in contrast (contrast decrements); respectively,

a total of 125 and 116 distorted images. The TID2013 and CSIQ databases contain, respectively, 25 and 30 reference images. These databases use a functional lookup table in the intensity component for decreasing or increasing (only TID2013) the contrast in the images. The TID2013 database uses 5 different distortion levels per reference (25 references $\times$ 5 distortion levels = 125 distorted images) and the CSIQ database has available 4 distortion levels for 26 reference images and 3 distortion levels for 4 reference images (26 references $\times$ 4 distortion levels + 4 references $\times$ 3 distortion levels = 116). Each distorted image has a subjective score for comparing the performance between fidelity measures. The subjective scores are expressed in terms of Differential Mean Opinion Scores (DMOS) for CSIQ and Mean Opinion Scores (MOS) for TID2013.

The MOS values from TID2013 were collected using a methodology known in psychophysics as two alternative forced choice (2AFC) match to sample (Ponomarenko et al., 2015). In 2AFC three images are displayed (the reference and two distorted images) and an observer selects one of the two distorted images which they judge as more similar to the reference. That is, human observers are asked to select among two images the image that perceptually differs less from a reference (Kingdom and Prins, 2010b). Thus, the evaluation is made in terms of the presented current stimuli. Since the 2AFC was made within the contrast subset of the TID2013, the MOS scores designated to that subset are a measure of the contrast difference with respect to the reference image perceived by the observers.

The DMOS values from CSIQ database were collected based on a linear displacement of the images. That is, all of the distorted versions of an original image were viewed simultaneously on a monitor array and placed in relation to one another according to the perceived quality difference (Larson and Chandler, 2010). The images were sorted by the observers according to the perceived differences with respect to the reference. Analogous to TID2013, since the rating was made within the contrast subset of the CSIQ, the DMOS scores designated to that subset are a measure of the contrast difference with respect to the reference image perceived by the observers.

The aforementioned test data descriptions show that TID2013 and CSIQ provide measurements of the perceived image differences and by isolating the contrast subset of each database, the (D)MOS scores become the measurements of the perceived contrast changes between the reference and the test samples. Therefore, the databases are appropriate to test the performance of the image contrast ratio measures in predicting the perceived contrast changes typically perceived and reported by a human observer. For further information about TID2013 and CSIQ databases, cf. (Ponomarenko et al., 2015; Larson and Chandler, 2010).

4.4.2 Performance comparison

We compare our proposed measure to the state-of-the-art measures listed in Table 4.1. Each of these methods is also evaluated by using the evaluation

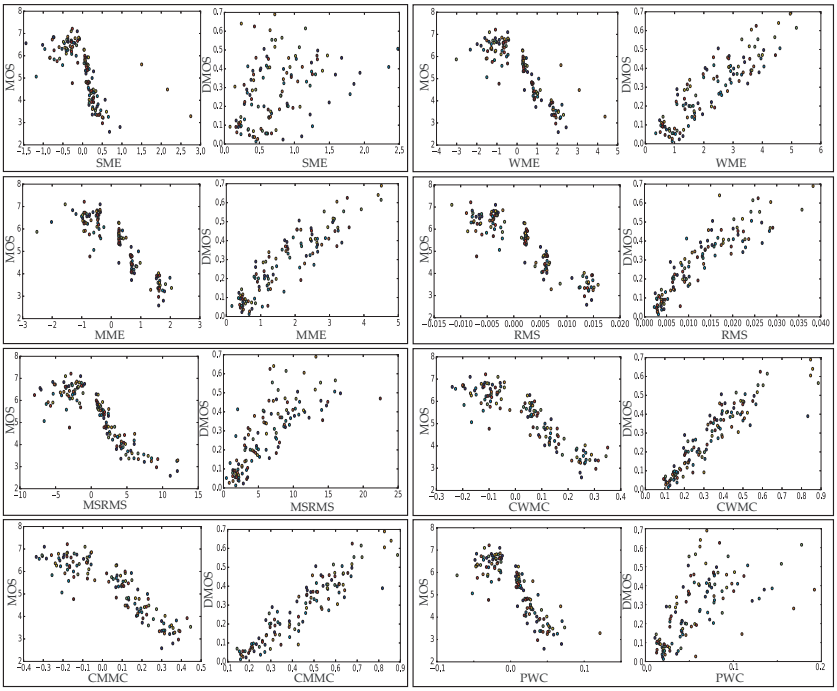

Figure 4.11: Scatter plots of the considered image contrast ratio measures and the subjective scores of TID2013 and CSIQ, in each box left and right, respectively.

methodology explained in Section 2.3.

Figure 4.11 shows the scatter plots of the considered image contrast ratio measures and the subjective scores of TID2013 and CSIQ. Note that, TID2013 data has negative contrast ratio differences (left scatter plot in each box). This is because as mentioned in the data set description, TID2013 includes not only contrast decrements but also increments. In the case of contrast increments, the contrast ratio in the reference image is lower than in the test image. The scatter plot for SME on CSIQ data shows the poor performance of this method in data with diverse image content, i.e., reference images covering a wide range of SA values as in CSIQ data (see Figure 4.9). MSRMS achieves a better performance than SME by using multi-scale and $RMSC_1$ formula. However, as we said earlier the $RMSC_1$ formula approximates contrast ratio by measuring the variability of the pixel values with respect to the central pixel which is unreliable on images with high spatial activity. Additionally, the scatter plots for SME and MSRMS on TID2013 show the poor performance of these contrast ratio measures in predicting image differences due to contrast increments.

WME and MME produce better results on CSIQ data. However, they also show a lower performance in predicting the image differences due to contrast increments of the TID2013 database. RMS (note that $RMSC_2$ takes into account

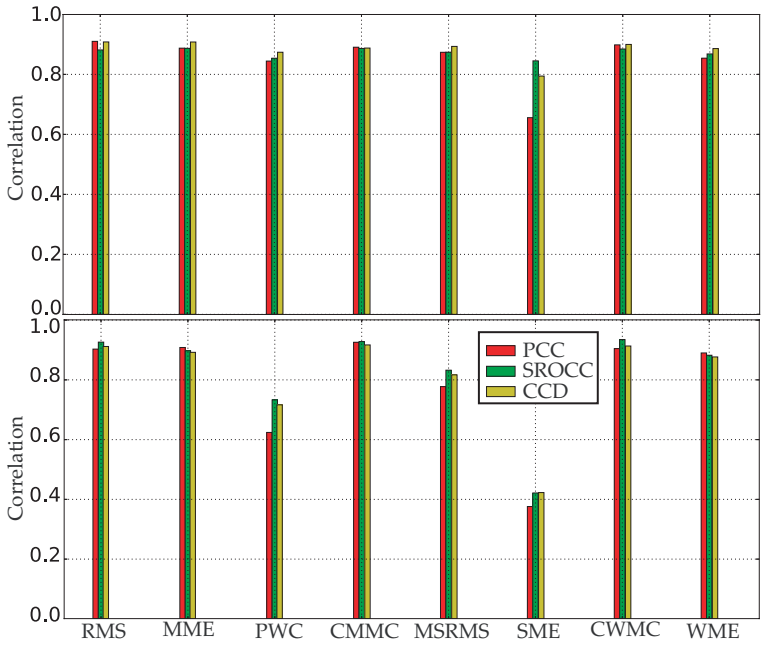

Figure 4.12: Performance of the considered image contrast ratio measures appraised on the (top) TID2013 and (bottom) CSIQ test sets.

the local average for the local ratio computation), CWMC and CMMC achieve the highest performance of the considered contrast ratio measures. We believe that this good performance is due to the fact that these measures are based on the local image content. However, RMS assumes that the central pixel is the foreground and the average intensity is the background which is not the case for edges making this measure undesirable for images with many perceptually significant edges as in CSIQ data. The PWC shows a very poor performance in CSIQ database. We attribute this low performance to the filter and parameter selection in the DWT (in this work the Haar wavelet and 4 decomposition levels [maximum allowed by the tested image sizes]) because it is known that the DWT provides different information about the edges depending on the filter and parameter selection (Provenzi and Caselles, 2014). That is, different filters and parameters could lead to different performances because the image content is characterized differently for each filter and/or selected parameter. However, that study is out of the scope of this thesis.

Figure 4.12 shows the performance of the considered image contrast ratio measures. Note the poor performance of SME measure in CSIQ database compared with TID2013. We attribute this low performance (correlation with human scores lower than 45% in CSIQ subset) to the fact TID2013 images possess little texture and spatial information compared with CSIQ. Thus, the

Table 4.2: Percentage increase of the performance appraised on TID2013 and CSIQ test sets of the proposed contrast ratio measure (CMMC and CWMC) compared with the state-of-the-art techniques (SME, WME, MME, RMS, MSRMS and PWC). The negative sign stands for a percentage decrease.

		TID2013						CSIQ					
		MME	MSRM	PWC	RMS	SME	WME	MME	MSRM	PWC	RMS	SME	WME
CMMC	PCC	1	5	15	-3	81	12	7	56	122	9	311	14
	SROCC	0	4	10	1	13	6	12	37	75	1	265	18
	CCD	0	0	5	-3	30	1	9	36	74	2	248	15
CWMC	PCC	3	8	18	-4	86	15	0	44	104	1	278	5
	SROCC	0	4	10	1	13	5	16	41	81	3	277	22
	CCD	-3	2	10	-3	35	5	8	34	71	1	243	13

SME measure is more suitable for TID2013 images. However, in highly textured images like CSIQ this method fails which is a big disadvantage given that natural images possess inherent texture features. WME and MME measures achieve overall a better performance than SME by using the difference between maximum and minimum pixel intensities. Although these measures outperform SME in terms of correlations with subjective scores, they still can be improved because these measures do not take into account the local image content. By computing contrast using the algorithm of Section 4.3, CWMC and CMMC outperform SME, WME and MME in both databases. Table 4.2 shows the percentage increase of the proposed method compared with the other state-of-the-art measures based on the correlation coefficients shown in Figure 4.12 after applying the Fisher's z transform.

As we stated earlier, the MSRMS is a measure of pixel variability and not a direct measure of contrast ratio. Therefore, the MSRMS fails in images with high texture like in CSIQ subset. Finally, CWMC and CMMC perform equally well as RMS in TID2013 but they perform better in the CSIQ subset. The results suggest that the wider the range of content available in the image set (extent of image details and texture), the more difficult the contrast change evaluation. Additionally, the results show that unlike the other tested methods, CWMC and CMMC are able to handle a wide range of image content (variety of texture and spatial activity). This is because CWMC and CMMC are based on image content relevant for perceived contrast ratio computation (see Section 4.3.1). Our experimental results also show that our method performs better than the current state-of-the-art methods (PCC, SROCC and CCD >90%).

Figure 4.13 shows the performance of the considered contrast ratio measure appraised on (top) TID2013 and (bottom) CSIQ per individual reference. That is, for each individual reference image, we assess the correlation between the contrast ratio measures and the subjective scores over all distortion levels. Particularly, the Figure shows a box plot which is a graphical representation of the 25 and 30 PCC, SROCC and CCD of each subset of data corresponding to the 25 and 30 reference images for the TID2013 and CSIQ databases, respectively. This method is very useful to detect if the tested measures are able to estimate the perceived differences across the range of distortion levels under the same

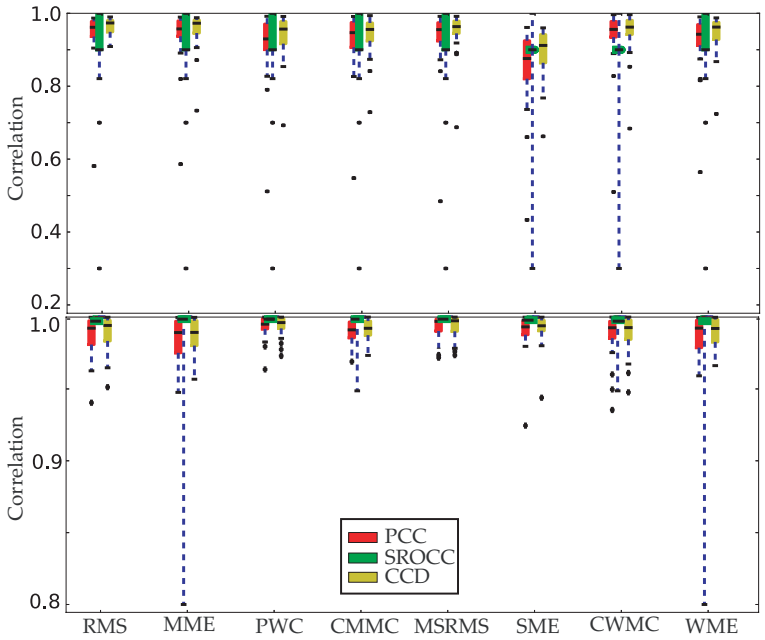

Figure 4.13: Performance of the considered image contrast ratio measures appraised on the (top) TID2013 and (bottom) CSIQ per individual reference image. The box plot was created using the (top) 25 and (bottom) 30 PCC, SROCC and CCD (one for each reference image) between a given contrast ratio difference measure and the corresponding subjective scores, respectively.

image content.

For instance, in the CSQI database all the methods perform with a CCD >95% (see Figure 4.13[bottom]) when the correlation is assessed between the contrast ratio measures and the subjective scores over all distortion levels for each individual reference image. This indicates that the contrast ratio measures are able to handle well the levels of distortions in CSQI database. However, SME, WME, MME, and MSRMS have a global performance with CCD <90% in the same subset (see Figure 4.12[bottom]). Thus, these contrast ratio measures can handle the levels of distortion included in the CSIQ data but they are not able to handle the wide range of different content included in the CSQI database (correlation assessed between the contrast ratio measures and the subjective scores over all distortion levels and reference images).

Figure 4.13(top) shows wider boxes than in Figure 4.13(bottom). This indicates that the contrast ratio measures are not able to handle the range of distortions for specific types of content (reference image) in the TID2013 database. We attribute this to the wider range of distortion levels included in TID2013 data (contrast increments and decrements) resulting in lower per-

Table 4.3: Table of dose levels, global contrast ratio measure values and subjective scores of the two static anthropomorphic chest phantoms.

Dose [μGy/s]	2613	1973	1302	978	643	308
MOS	77	73	72	74	52	33
RMS	0.02339	0.02592	0.03030	0.03340	0.03815	0.04556
CWMC	0.78702	0.76258	0.77196	0.77515	0.73204	0.69093
CMMC	0.72808	0.68138	0.68828	0.70416	0.64906	0.61174
Dose [μGy/s]	9330	6700	4980	4140	3320	2475
MOS	75	69	66	63	52	46
RMS	0.03382	0.03749	0.04074	0.04263	0.04475	0.04778
CWMC	0.72545	0.74521	0.74417	0.74974	0.75063	0.73490
CMMC	0.63531	0.68218	0.67944	0.69205	0.69871	0.68358

formances when the correlation is assessed in each individual reference image. Nevertheless, the global performance achieved by the contrast ratio measures in TID2013 is higher compared to their global performance in the CSIQ data. This confirms the fact that the content in the CSIQ is more challenging for computing contrast ratio than the content presented in TID2013 images. In any case, this does not apply for CWMC, CMMC and RMS where the global performance is consistent both within the database and between databases.

We attribute the good performance of the proposed contrast ratio measure to the fact that the proposed measure characterizes the local distribution of pixel values before the computation of the (local) contrast ratio. Therefore, the proposed method can predict the perceived differences due to contrast decrements/increments reported by human observers better than the other state-of-the-art methods tested in this work. Additionally, since there is no notable differences in the performance achieved by the proposed contrast ratio measure in the two databases, we can conclude that the proposed measure is able to handle successfully low and high textured images (a wide variety of content).

4.4.3 Measuring contrast ratio changes in interventional x-ray

Interventional X-ray refers to a range of techniques which rely on the use of radiological image guidance to precisely target therapy (BSIR, 2016). For instance, contrast-based fluoroscopy is one of the most frequently used diagnostic/interventional techniques in cardiology, e.g., dynamic x-ray imaging used to diagnose/treat cardiac conditions (Gislason et al., 2010). Thus, the number of x-ray fluoroscopies has increased over the past years increasing the radiation dose on patients and radiation-induced problems, for example, transient and permanent skin damage. Therefore, this kind of procedures has set a maximum limit on fluoroscopy patient exposure rates (measured as Entrance Skin Dose rate [ESD] $\approx$ 1500 microgray per second [μGy/s]) (US-FDA, 2005).

Thus, in this specific medical use case the goal is to identify the minimum

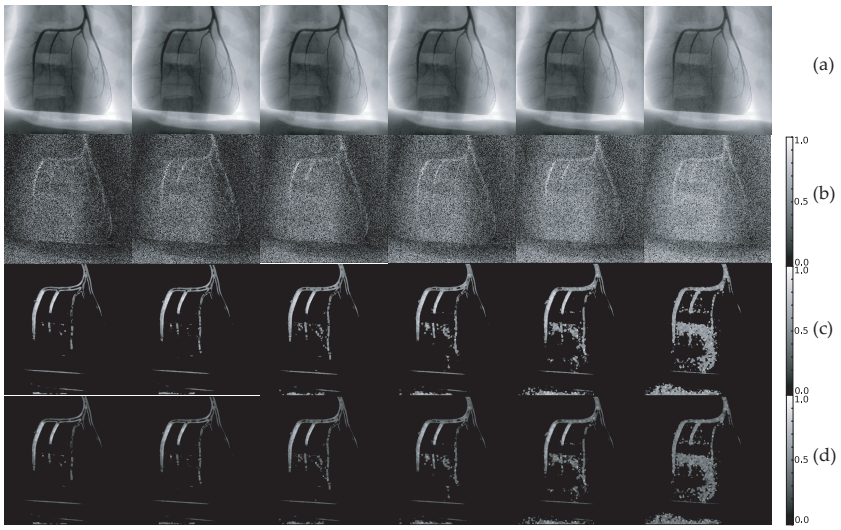

Figure 4.14: From left to right chest phantom scanned at 2613 μGy/s, 1973 μGy/s, 1302 μGy/s, 978 μGy/s, 643 μGy/s, 308 μGy/s, respectively. From top to bottom (a) input image, (b) *contrast ratio maps* of RMS, (c) CWMC and (d) CMMC.

radiation dose that results in the clinically relevant image contrast ratio and to acquire the images at that exact radiation dose (Gislason et al., 2010; Kumcu et al., 2015a). Here, the contrast ratio measure is used to quantify the contrast difference between the current and the reference image (image where the diagnostically relevant details [e.g., the coronary tree] are presented under "ideal" detectability conditions) and then from this contrast ratio difference we can predict the difference in visibility of the clinically relevant structures in the image.

In this Section, we evaluate the performance of the proposed methodology in mimicking the subjective ratings reported by cardiologist/radiologist in interventional X-ray images. We use a static anthropomorphic chest phantom scanned with and without a 10 cm polymethyl methacrylate plate to simulate standard and large chest thickness, respectively, at six dose levels. The dose levels, global contrast measure values and subjective scores of the two static anthropomorphic chest phantoms are shown in Table 4.3. The images were evaluated by 4 interventionalists (cardiologist/radiologist) from Ghent University Hospital in Belgium resulting in a Mean Opinion Score per image (Kumcu et al., 2015b). The interventionalists rated the similarity of each pair of images using a continuous scale from 0 (completely different) to 100 (exactly the same). See Ref (Kumcu et al., 2015b) and Section A.3 for further details about this database.

In the following paragraphs we compare the best performing contrast ratio measures from the experiments reported in Section 4.4.2: the root mean

Figure 4.15: From left to right chest phantom with 10 cm polymethyl methacrylate scanned at 9330 μGy/s, 6700 μGy/s, 4980 μGy/s, 4140 μGy/s, 3320 μGy/s, 2475 μGy/s, respectively. From top to bottom (a) input image, *contrast ratio maps* of (b) RMS, (c) CWMC and (d) CMMC.

squared measure of enhancement (RMS) (Panetta et al., 2013), content-aware Weber's and Michelson's measure of contrast (CWMC and CMMC, respectively). Figures 4.14 and 4.15 show the *contrast ratio maps* of the block-wise computations of RMS, CWMC and CMMC on the two static anthropomorphic chest phantoms. By comparing the contrast ratio values around the edges of the *contrast ratio maps* on CWMC and CMMC (Figures 4.14 and 4.15), we can see that these values come closer to the contrast ratio values of the noise (noise visibility) with the dose reduction. This is an indication of losing the relevant details (edges) due to low contrast. Unlike CWMC and CMMC, the RMS *contrast ratio map* does not show such a behavior, making it a less attractive method for applications where local content needs to be analyzed such as in medical imaging.

Figure 4.16 shows the scatter plots of the considered image contrast ratio measures and dose level (left) and the subjective scores with its standard deviation (right) for the two static anthropomorphic chest phantoms. The scatter plot for RMS shows that in the large chest there are no statistically significant differences between dose levels in terms of MOS, i.e., there are no perceived differences between the images. However, the RMS displays a wide range of values for the same MOS level. Thus, we confirm once more that RMS is not a direct measure of contrast ratio and therefore an increase or decrease of its value may not necessarily reflect a significant change in perceived contrast. This is important because it implies that increasing the dose level does not

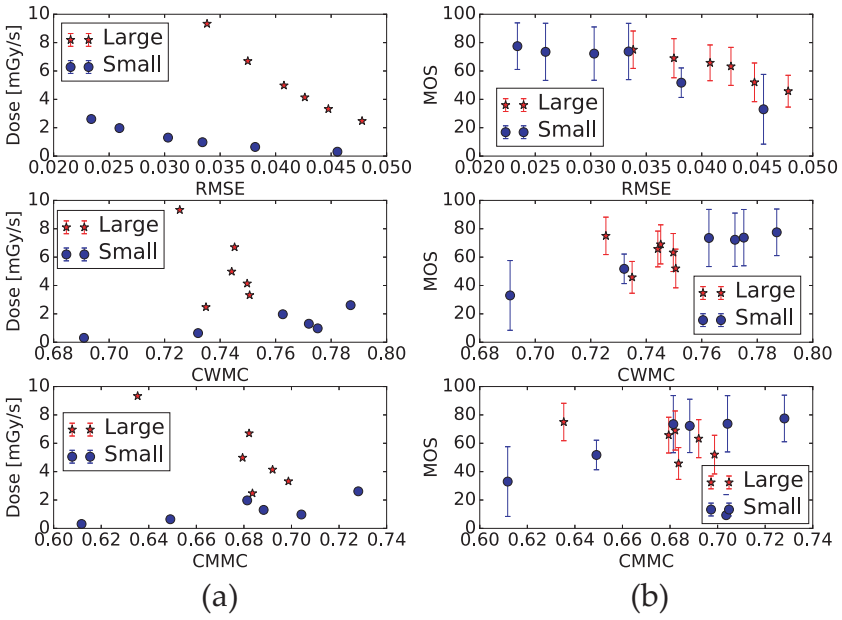

Figure 4.16: Scatter plots of the considered image contrast ratio measures and (left) dose level and (right) the subjective scores of the two static anthropomorphic chest phantoms.

necessarily result in contrast changes, i.e., in perceivable differences between the images.

For example, the first dose level for the large phantom is 2475 μGy/s which is 1.65 times higher than the recommended maximum 1500 μGy/s. Therefore, the first dose level should have enough visibility for the interventionalists and increasing the dose level would not improve the visibility around the structures of interest (edges). For the small phantom, the dose levels are below the recommended maximum dose value. In such a case, it is possible to see that there are statistical significant differences between dose levels in terms of MOS. That is, the interventionalists perceived image differences between the dose levels. Therefore, from the lowest dose level there is still room for increasing the contrast (induce image differences). In either case, CWMC and CMMC are able to predict these changes in contrast (perceived image differences) except at the highest dose level in the large phantom which is in general a non-practical or unrealistic situation due to the excessive radiation dose (6.22 times the recommended maximum dose level). This is a good starting point for designing an automatic dose control system based on the contrast ratio. For example, to select the appropriate dose levels by considering the size of the patient and the desired contrast level with respect to previously recorded standard image samples (reference) where the diagnostically relevant details are presented under

"ideal" detectability conditions.

4.5 Conclusions

We proposed a measure to compute contrast ratio in local image patches. The main novelty is in the use of bimodal histograms to compute the local contrast ratio. Next, we performed an extensive experimental evaluation based on a total of 6 image contrast ratio measures, each tested on 241 distorted images (references altered for contrast). We have tested Weber's and Michelson's contrast ratio formulas in the proposed local contrast estimation to simulate the case where a small structure of interest is present on an uniform background or a square-wave grating of one cycle, respectively. Moreover, we have tested our methodology on predicting the perceived differences reported by medical experts for interventional X-ray images. To stimulate further experimentation, we made all the tested methods freely available as a plugin on the iFAS software tool.

The results show that our method is able to accurately predict the perceived image differences due to contrast decrements/increments. This could be due to the fact that the proposed method uses the local distribution of pixel values for computing the contrast ratio instead of the maximum and minimum pixel intensities which in general leads to errors in highly textured areas. Additionally, we have shown the major advantages of our method in interventional X-ray fidelity assessment.

Percentile pooling over the computed *contrast ratio maps* was tested. We found that the 75% threshold level of the harmonic mean is the best performing threshold for predicting fidelity changes due to contrast increments/decrements based on humans scores in TID2013 and CSIQ databases. Additionally, the impact of the patch size and overlap parameter in the proposed methodology were demonstrated. The results suggest that the patch size does not have a big impact on the performance of the proposed measure. Overall, given the test images in our work, we found that a patch size of 15×15 produces slightly better predicted contrast ratio values than the other patch sizes tested in this work. The patch size can be further investigated for those applications where it is required to use multiple image resolutions, e.g., if the measure needs to be computed on 1080p images or 4K images. This problem could be potentially solved by using a multi-resolution approach. That is, by analyzing the contrast ratio values on different image resolutions using image sub-sampling. Additionally, we found that decreasing the overlap parameter does not have a major impact on the performance but it reduces considerably the computational time.

The contributions reported in this Chapter resulted in two international conference proceedings (Ortiz-Jaramillo et al., 2015b,a), and one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2018a).

5

Assessment of appearance changes in texture

5.1 Introduction

The evaluation of appearance parameters to determine lifetime of textile products (appearance retention) is one of the main concerns for manufacturers. Appearance retention refers to the capability of the materials in retaining the original appearance under common conditions of daily use (Orjuela-Vargas, 2012). The evaluation of appearance parameters is typically conducted by visual inspection of parameters such as pilling resistance, abrasion resistance, shrinkage, wrinkles, drape, and color, according to the product under inspection. The visual inspection is typically carried out by human experts (Hu, 2008). In general, the visual inspection covers identification of defective areas and the evaluation of appearance changes in the surface of the textile material. However, visual inspection has shown to be unreliable and costly (Waegemana et al., 2008). Therefore, computerized automatic visual inspection, mostly using texture analysis algorithms, has been used to alleviate those problems.

In the evaluation of surface appearance in textile materials, the most important visual parameters are texture and color (Smeulders et al., 2000; Hiremath and Pujari, 2007; Aptoula and Lefevre, 2011). Figure 5.1 shows examples of standard grading level sets of transitional appearance changes of textile material surfaces. Wrinkling assessment (Figure 5.1(a)) is used to determine the ability of the product of retaining or recovering a smooth surface appearance after repeated use (Na and Pourdeyhimi, 1995; Zhifeng et al., 2003). In pilling assessment, the surface area of entangled fibers that remain attached to the original surface is estimated (Mendes et al., 2010). In pile surface assessment, the surface structure of the used material is compared to the corresponding original surface (Orjuela-Vargas, 2012). Overall, the evaluation of appearance changes consists in visually identifying categorized deviations from a reference sample. Therefore, fidelity assessment plays an important role in the development of automatic visual inspection systems for measuring appearance changes.

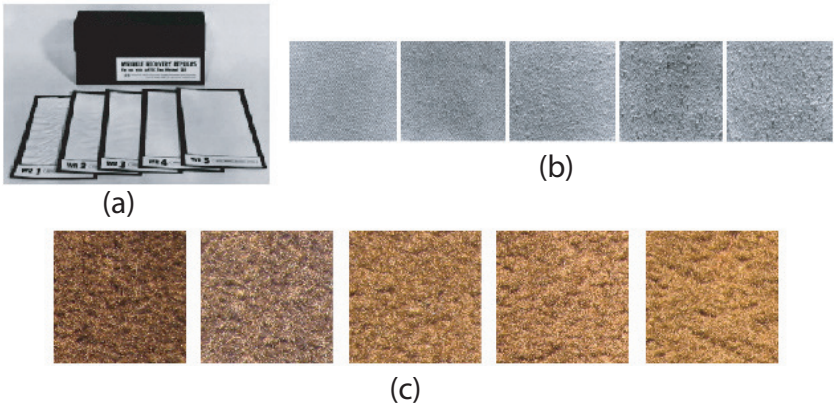

Figure 5.1: Examples of standard grade level sets of transitional appearance changes of textile material surfaces. (a) wrinkling, (b) pilling, (c) pile surface assessment.

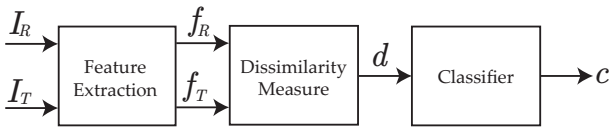

Figure 5.2: General procedure for fidelity assessment of textiles (evaluation of appearance changes in texture).

In this field, fidelity assessment is used to evaluate visual deviations of texture and color between a textile sample and a reference (new textile). As stated in previous paragraphs, the fidelity assessment process is still carried out by human visual inspection, which is laborious, repetitive, exhausting, insufficient and costly. Also, the human assessment lacks reproducibility and results in inconsistencies between judges (Waegemana et al., 2008). Therefore, researchers have proposed to automate the fidelity assessment of textile surfaces using texture analysis approaches (Siew et al., 1988; Davies and Hall, 1999; Kang et al., 2005; Xie, 2008; Orjuela-Vargas et al., 2010; Xin et al., 2011), i.e., by evaluating appearance changes in texture. Automatic fidelity assessment of textiles includes applications like roughness measurement, wrinkling evaluation, seam puckering assessment, pilling assessment and evaluation of appearance changes in floor coverings. Additionally, texture analysis approaches have been successfully applied in other tasks such as image segmentation, texture classification, defect detection, texture synthesis and estimation of image deformations (Mao and Jain, 1992; Joshi et al., 2009; Abbadeni, 2010; Elunai et al., 2010).

Figure 5.2 shows a general procedure for the evaluation of appearance changes in texture using image processing (Siew et al., 1988; Waegemana et al., 2008; Orjuela-Vargas et al., 2008, 2010). Features are extracted independently

from the picture of the reference (I_R) and the test (I_T) texture samples, termed $\mathbf{f}_R$ (vector of features extracted from the reference sample) and $\mathbf{f}_T$ (vector of features extracted from the test sample), respectively. Afterwards, a dissimilarity measure between the extracted features is computed, i.e., $d(\mathbf{f}_R, \mathbf{f}_T)$ where the function $d(\cdot)$ is an appropriate distance measure. Finally, according to the value given by the distance measure, a label c representing the wear degree of the physical sample is established (Orjuela-Vargas et al., 2010).

Among the most common used techniques (extracted features) in evaluating appearance changes of texture are the following: co–occurrence matrices (Siew et al., 1988; Mori and Komiyama, 2002; Mak and Li, 2008), filter bank decomposition (Saint-Marc et al., 1991; Wood, 1993; Palmer et al., 2009), granulometry (Aibara et al., 1999; Davies and Hall, 1999; Xin et al., 2011), grey level differences (Pourdeyhimi et al., 1994a), grey value histogram analysis (Jose et al., 1986), power spectrum (Wang and Wood, 1994; Xu, 1997; Jensen and Carstensen, 2002; Choi et al., 2009), the Radon transform (Mohri et al., 2005), spatial grey level dependences (Pourdeyhimi et al., 1994b), wavelet analysis (Militký and Bleša, 2008; Kang et al., 2005; Sun et al., 2011), Gaussian models (Abril et al., 1998), local binary patterns (Orjuela-Vargas, 2012) and the Wigner distribution (Cristobal and Hormigo, 1999).

The aim of this chapter is to review and evaluate texture analysis descriptors for automatic evaluation of appearance changes in texture. Additionally, we discuss the impact of the parameter selection in the evaluated texture analysis techniques. To evaluate the reviewed techniques, we consider the degradation appearing on the surface of textile floor coverings. Our findings show that DWT, Eig, FFT and Gb provide good descriptors for assessing degradation in textile floor coverings exhibiting strong correlations with the assessment of human experts. Particularly, in the evaluation of floor coverings with cut-pile and loop-pile types, i.e., texture surfaces without complex patterns. Therefore, these features can be used as starting point in applications involving the assessment of appearance changes in texture, as well as a basis for the development of new methods.

This Chapter is organized as follows: in Section 5.2, current texture approaches commonly used in the evaluation of surface changes in industrial web materials are reviewed. Thereafter, in Section 5.3, we discuss the results obtained in our particular case of study. Finally, in Section 5.4 conclusions and future work are drawn.

5.2 Image texture analysis

Most of natural and artificial surfaces exhibit texture. The characterization of such surfaces like plants, fabrics, minerals, skin is performed by using texture analysis. Texture analysis is one of the main concerns in computer vision due to its wide range of applications such as remote sensing, medical diagnosis, fidelity assessment and quality control (Chen, 1995; Tuceryan and Jain, 1998; Zhang and Tan, 2002).

Numerous methods have been proposed to deal with different visual texture inspection tasks (Haralick, 1979; Ojala et al., 1996; Tuceryan and Jain, 1998; Randen and Husoy, 1999; Zhang and Tan, 2002; Singh and Singh, 2002; Popescu et al., 2007; Xie, 2008; Mazher and Ali, 2011; Orjuela-Vargas, 2012). Within those tasks it is possible to find a wide range of definitions concerning to the question *what is texture?* For instance,

(Tamura et al., 1978) write that *"we may regard texture as what constitutes a macroscopic region. Its structure is simply attributed to the repetitive patterns in which elements or primitives are arranged according to a placement rule"*.

(Cambridge dictionary, 2016) defines texture as *"the quality of something that can be decided by touch; the degree to which something is rough or smooth, or soft or hard"*.

(Zhang et al., 2010b) refer to texture as a *"measure of the variation of the intensity of a surface, quantifying properties such as smoothness, coarseness and regularity"*.

According to (Nixon and Aguado, 2002): *"texture is actually a very nebulous concept, often attributed to human perception, as either the feel or the appearance of (woven) fabric"*.

The above definitions show that texture is easily perceived by humans but definitions are given according to the application and there is not a unique definition for the word texture. Particularly, in image processing, texture is defined as a *"function of the spatial variation in pixel intensities"* (Tuceryan and Jain, 1998), i.e., texture analysis attempts to provide information about the spatial arrangement of intensities in an image. Thus, the purpose of texture description is to derive some measurements that can be used for identifying certain useful characteristics for a texture classification task.

According to (Tuceryan and Jain, 1998; Zhang and Tan, 2002; Xie, 2008) texture descriptors can be grouped into four categories: statistical features, structural features, signal processing based features and model based features. In the rest of this Chapter the most commonly used techniques for texture analysis using image processing are explained and evaluated in the assessment of appearance changes in textiles. Note that the texture analysis techniques discussed in this Chapter use gray scale images to compute the texture descriptors.

5.2.1 Statistical features

These techniques are used to measure the spatial distribution of pixel values at specific positions (Xie, 2008) and are applied in tasks such as texture analysis, image segmentation, texture classification, defect detection, wear evaluation (Haralick, 1979; Swain and Ballard, 1990; Orjuela-Vargas et al., 2010). The most popular techniques in this category are described in the following paragraphs.

Autocorrelation function

Two different textures can be distinguished by evaluating differences in their regularity or fineness presented in the image. One way of measuring this kind of differences is by computing separately the autocorrelation function in both textures (Petrou and Sevilla, 2006). Next the resulting autocorrelation functions are described using a parametric model. Afterwards, the model parameters are compared to obtain the texture differences. The autocorrelation function has been used in several applications including fabric analysis, macro-texture analysis, estimation of deformation (Heilbronner, 1992; Torabi et al., 2008; Elunai et al., 2010). The autocorrelation function of a given image I of $M \times N$ pixels is defined as:

$$\rho\left(I\right) = \frac{\mathcal{F}^{-1}\left\{\mathcal{F}\left\{I\right\}\mathcal{F}\left\{I^{*}\right\}\right\}}{e}, \tag{5.1}$$

where $e = \sum_{i=1}^{M}\sum_{j=1}^{N} I^{2}(i,j)$ is the energy of I, I^{*} is the complex conjugate of I, $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ are the direct and inverse discrete Fourier transform, respectively. Practical algorithms use $\rho\left(I\right)$ to characterize the texture under analysis (Petrou and Sevilla, 2006).

Co-occurrence matrix

This method is one of the most well-known texture analysis techniques, especially, in surface flaw detection (Siew et al., 1988; Orjuela-Vargas et al., 2008). The co-occurrence matrix is a matrix of frequencies $P(g_1, g_2|(x,y))$, where g_1 and g_2 are pixel values of two disjoint pixels separated by a displacement (x,y). The number of occurrences of g_1 and g_2, separated by (x,y), contributes to the (g_1,g_2)th entry in the matrix of frequencies P. Those matrices characterize the distribution of co-occurring pixel values giving the displacement (x,y) (Tuceryan and Jain, 1998; Randen and Husoy, 1999).

Local Binary Patterns (LBP)

This method computes relative intensity relations between the pixels in a small neighborhood compared to the central pixel (g_c). The LBP is a widely used texture descriptor including a wide range of its extensions (Maenpaa, 2003).

The LBP code for an eight pixel neighborhood of Figure 5.3(a) is computed as

$$\mathrm{LBP} = \sum_{k=0}^{7} T\left(g_k - g_c\right) 2^{k},$$

where

$$T(x) = \begin{cases} 0, & \text{if} \quad x < \epsilon \\ 1, & \text{if} \quad x \geq \epsilon \end{cases},$$

for a given small value ϵ. g_c is the central pixel and g_k for $k = 0, \ldots, 7$ are the corresponding neighboring pixels of Figure 5.3(a). After computing texture codes per pixel, these codes are grouped into a histogram to characterize the texture of an image.

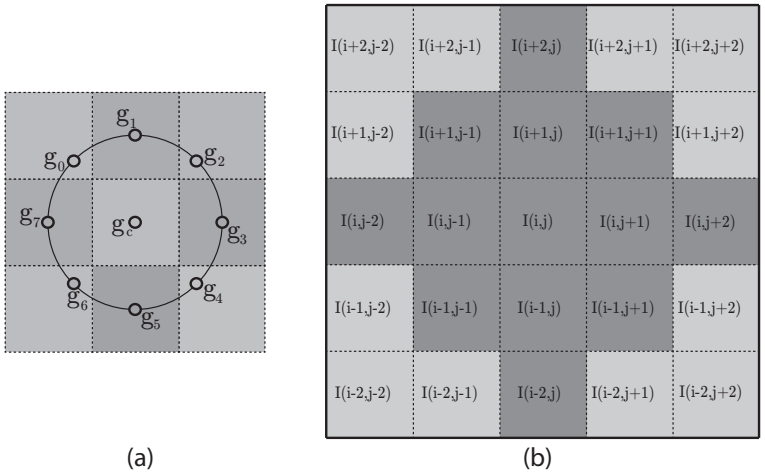

Figure 5.3: (a) Pixel set distributed within a circularly symmetric neighborhood. Here, an eight pixel circularly symmetric neighborhood. (b) Third order Markov neighborhood within a 5×5 window.

5.2.2 Model based features

Generally, this type of features approximates the texture under analysis using parametric models. Then, the estimated model parameters are used as texture features (Zhang and Tan, 2002). Also, those parameters are used very often to synthesize textures (Tuceryan and Jain, 1998). The most popular features in this category are described in the following paragraphs.

Autoregressive models

The autoregressive (AR) models use linear models to describe the relationship between neighboring image pixels (Randen and Husoy, 1999; Zhang and Tan, 2002; Xie, 2008; Joshi et al., 2009). The AR models have been successfully applied in several tasks such as texture synthesis (Mao and Jain, 1992), texture segmentation (Joshi et al., 2009) and texture classification (Abbadeni, 2010). A two-dimensional AR model is defined as the linear combination of the surrounding neighbors of a central point (Deguchi, 1986). The AR model for the central pixel (g_c) of the Figure 5.3(a) is given by

$$g_c \approx \sum_{k=0}^{7} a_k g_k, \tag{5.2}$$

where a_k for $k = 0, \ldots, 7$ are the parameters of the linear model to be estimated and g_k for $k = 0, \ldots, 7$ are the corresponding neighboring pixels of the central pixel g_c. Particularly, this method is called circular autoregressive

model (Kashyap and Chellappa, 1983). Equation (5.2) can be generalized for every pixel in the image as

$$I(i,j) \approx \sum_{x=-1}^{1} \sum_{y=-1}^{1} a(x,y) I(i+x, j+y).$$

The set of parameters $a(x,y)$ is estimated by means of linear regression and used to characterize the texture.

Gaussian Markov random field (GMRF)

Markov random fields (MRFs) are a probabilistic representation of all interactions between pixels values within a local neighborhood. In other words, it describes the global distribution of pixels values in terms of local neighborhood interactions (Cross and Jain, 1983; Krishnamachari and Chellappa, 1997). MRFs have been very popular for modeling images and applied in several tasks such as texture synthesis (Paget and Longstaff, 1995) and classification (Chen and Huang, 1993), image segmentation (Yang and Jiang, 2003), restoration (Babacan et al., 2008) and compression (Krishnamoorthi and Seetharaman, 2007). In (Cross and Jain, 1983) the binomial model is used to produce blurry, sharp, line-like, and blob-like textures. The results showed that the synthetic textures closely resembled their real counterparts (Cross and Jain, 1983). The GMRF have been shown to be accurate in applications involving texture segmentation (Krishnamachari and Chellappa, 1997). Additionally, the parameters of a GMRF, unlike to other MRF extensions, can be computed efficiently (Grath, 2003). Note that these type of models can be also categorized as well as a statistical based feature because of its use of statistics for describing local patches. However, they are mostly categorized within the model based features in the texture analysis field. The GMRF model for the set of pixels in Figure 5.3(b) is defined by the following formula:

$$p(I(i,j)|I(i+x,j+y),(x,y) \in [-2,2]) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(I(i,j) - \sum_{l=1}^{6} \alpha_l S_{i,j,l}\right)^2\right),$$

where $S_{i,j,l}$ for $l = 1, \ldots, 6$ are defined as follows:

$$\begin{array}{rcl} S_{i,j,1} &=& I(i-1,j) + I(i+1,j) \\ S_{i,j,2} &=& I(i,j-1) + I(i,j+1) \\ S_{i,j,3} &=& I(i-2,j) + I(i+2,j) \\ S_{i,j,4} &=& I(i,j-2) + I(i,j+2) \\ S_{i,j,5} &=& I(i-1,j-1) + I(i+1,j+1) \\ S_{i,j,6} &=& I(i+1,j-1) + I(i-1,j+1) \end{array} \tag{5.3}$$

The seven parameters $\{\alpha_1, \ldots, \alpha_6, \sigma\}$ are used to characterize the texture under analysis.

5.2.3 Structural features

From a structural point of view, texture is characterized by texture primitives and the spatial arrangement of those primitives (Tuceryan and Jain, 1998; Zhang and Tan, 2002; Xie, 2008). Thus, texture primitives consider spatial basic structures and their placement as well as orientation features to characterize texture (Aptoula and Lefevre, 2011).

Granulometry

Mathematical morphology is the most well-known and used tool for extracting structural texture primitives (Aptoula and Lefevre, 2011). Those primitives are normally computed by using a pattern called structuring element (SE). The majority of morphological texture features depend on a set of morphological transformations, i.e., on applying successively morphological operations while increasing the SE size. This results in a set of images with less and less details characterizing structures through the scales. Particularly, when a set of morphological openings are applied, the resulting set of images is called granulometry. Similarly, anti-granulometry can be measured using a set of morphological closings instead of openings. The morphological opening operation is the dilation of the erosion of a set by a SE and the morphological closing is the erosion of the dilation of the same set using the same SE, where erosion and dilation are the two fundamental operations of mathematical morphology (Aptoula and Lefevre, 2011). While granulometry captures bright details on dark background, anti-granulometry focuses on dark details on bright background (Aptoula and Lefevre, 2011). Both sets characterize the granularity of the texture.

5.2.4 Signal processing based features

These approaches use spatial-frequency analysis to extract features. Most of the signal processing features are extracted by applying linear transformations, filtering or filter bank decomposition, followed by some energy computation (Tuceryan and Jain, 1998; Randen and Husoy, 1999; Xie, 2008). These features are derived from spatial, spatial-frequency and joint spatial/spatial-frequency domain. In this Section, we describe the most popular signal decomposition techniques used in texture analysis.

Power spectrum

The Fourier transform (FT) is a mathematical operation that decomposes a function into its constituent frequencies, also known as a frequency spectrum. The FT is used in a wide range of applications, such as image analysis, image filtering, image reconstruction and image compression (Lee and Chen, 2002; Petrou and Sevilla, 2006; Park et al., 2010; Kaur et al., 2011). The FT is useful when it is necessary to access to the geometric characteristics of the spatial

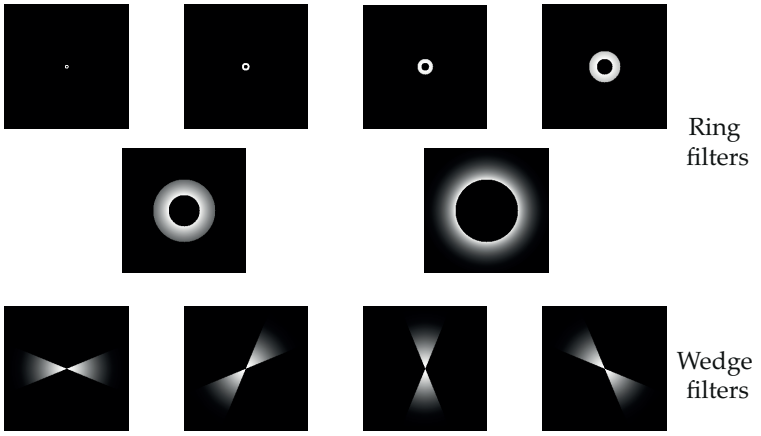

Figure 5.4: Set of ring and wedge filters in a dyadic configuration.

domain (Nixon and Aguado, 2002). The two-dimensional Fourier transform is defined as

$$\mathcal{F}\{I\}(u,v) = \sum_{i=0}^{M}\sum_{j=0}^{N} I(i,j)\exp\left(-2\pi\left(\frac{ui}{M}+\frac{vj}{N}\right)\right), \tag{5.4}$$

where u and v are the spatial frequencies across the rows and columns of the image. A set of wedge and ring filters have been suggested to discriminate texture in spatial-frequency and orientation (Weszka et al., 1976; Randen and Husoy, 1999).

The filters proposed by (Weszka et al., 1976) have shown to be accurate for image analysis (see Figure 5.4):

$$\begin{aligned} h_{r_1,r_2} &= \sum_{\substack{r_1^2 \le u^2+v^2 < r_2^2 \\ 0 \le (u,v) < (M,N)}} |\mathcal{F}\{I\}(u,v)|^2 \\ h_{\theta_1,\theta_2} &= \sum_{\substack{\theta_1 \le \tan\left(\frac{v}{u}\right) < \theta_2 \\ 0 < (u,v) < (M,N)}} |\mathcal{F}\{I\}(u,v)|^2 \end{aligned} \tag{5.5}$$

where h_{r_1,r_2}, h_{θ_1,θ_2} are ring and wedge filters, respectively. r_i and θ_i are the bounds of the filters. The filters are used to decompose the image in frequency sub-bands and descriptive statistics are computed on each sub-band to characterize the texture.

Eigenfilter

The ability to incorporate various spatial and spatial-frequency constraints make eigenfilters a useful tool for signal analysis and synthesis (Ade, 1983), as

well as in a variety of tasks such as image classification, image segmentation, defect detection (Ade, 1983; Tkacenko and Vaidyanathan, 2003; Monadjemi et al., 2004). The eigenfilters are computed as follows: first the image pixels are arranged into row entries of a matrix E of size $d \times W^2$, where d is the number of overlapping windows used to visit every pixel in the image using a window of size $W \times W$ (usually window size is 3×3). Afterwards, the covariance matrix is computed from the matrix E as $B = E^T E$. Then, after applying a singular value decomposition over the covariance matrix B, the obtained W^2 eigenvectors of such a decomposition are used as filters to perform a filter bank decomposition that results into a set of images. Usually, descriptive statistics are computed on this set of filtered images to characterize the texture.

Gabor filters

Gabor filtering allows the representation of images with optimal joint localization in the spatial-spatial/frequency domains (Jain and Farrokhnia, 1990; Manjunath and Ma, 1996; Randen and Husoy, 1999; Liu and Wang, 2003) characterizing the human vision system (Jain and Farrokhnia, 1990; Manjunath and Ma, 1996; Ortiz-Jaramillo et al., 2012). This makes this filtering approach very successful for the estimation of quality of compressed images/videos (Seshadrinathan and Bovik, 2010; Ortiz-Jaramillo et al., 2012). A two-dimensional Gabor function $g(x, y)$ and its Fourier transform $G(u, v)$ are defined as (Manjunath and Ma, 1996),

$$\begin{array}{rcl} g(x,y) & = & \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi j u_0 x\right) \\ G(u,v) & = & \exp\left(-\frac{1}{2}\left(\frac{(u-u_0)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right)\right) \end{array} , \tag{5.6}$$

where $\sigma_u = (2\pi\sigma_x)^{-1}$ and $\sigma_v = (2\pi\sigma_y)^{-1}$. Here, σ_u and σ_v characterize the band width of the Gabor filter centered at the point $(u_0, 0)$ in the spatial/frequency domain (u, v). From Equation (5.6), it is possible to generate a set of Gabor functions by appropriate dilations and rotations, i.e.,

$$\begin{array}{rcl} g_{m,n} & = & a^{-m} g(x', y'), \;\; a > 1, \;\; m, n \in \mathbb{Z} \;\; \text{with} \\ x' & = & a^{-m}(x\cos(\theta) + y\sin(\theta)), \;\; \text{and} \\ y' & = & a^{-m}(-x\sin(\theta) + y\cos(\theta)), \end{array}$$

where $\theta = n\pi/K$, with $n = 0, \ldots, K-1$, and K is the total number of orientations. Here a^{-m}, with $m = 0, \ldots, S-1$, is the scale parameter, where S is the number of scales.

Particularly, (Manjunath and Ma, 1996) proposed a strategy to reduce the redundancy presented in this filter bank decomposition. The strategy ensures that the responses of the filters in the spatial/frequency domain are tangent to each other (see Figure 5.5). Let U_l and U_h denote the lower and upper center frequencies of interest, respectively. Then, the filters are defined in terms of:

$$a = \frac{U_h}{U_l}^{\frac{1}{S-1}}$$

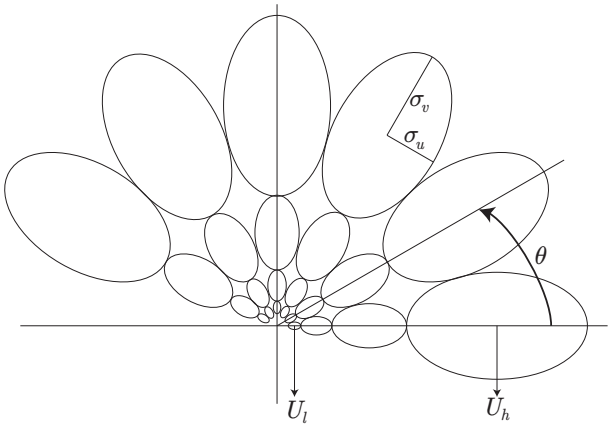

Figure 5.5: Set of Gabor filters in the spatial-frequency domain. The elliptical contours tangent to each other indicate the response of the Gabor functions.

$$\begin{aligned}\sigma_u &= \frac{(a-1)U_h}{(a+1)\sqrt{2\ln(2)}} \\ \sigma_v &= \left(\frac{\pi}{2K}\right)\left(U_h - 2\ln\left(\frac{2\sigma_u^2}{U_h}\right)\right)\left(2\ln(2) - \frac{(2\ln(2))^2\sigma_u^2}{U_h^2}\right)^{-\frac{1}{2}}\end{aligned}$$

where $u_0 = U_h$. As shown in Figure 5.5, the filters are rotated versions of the Equation (5.6), where σ_u and σ_v are related to a. This set of filters is used to decompose the image into frequency sub-bands. In general, descriptive statistics are computed on the resulting filtered images to characterize the texture.

Laplacian pyramid

The Laplacian pyramid is a multi-scale technique that has been used in applications such as image restoration (Xingmei et al., 2010), contrast enhancement (Dippel et al., 2002) and texture analysis (Chan et al., 2003; Vo et al., 2006; Yong et al., 2010). This technique is often used to characterize the image details at different image scales. The Laplacian pyramid consists of a sequence of differences between two consecutive levels of the Gaussian pyramid (Burt and Adelson, 1983), i.e., $I_{k+1} = I_k - \widetilde{I}_k$, where I_k is the image I at the kth level and $\widetilde{I}_k$ is the image I at the kth level after Gaussian filtering (see Figure 5.6). Descriptive statistics are computed on the resulting filtered images to characterize the texture.

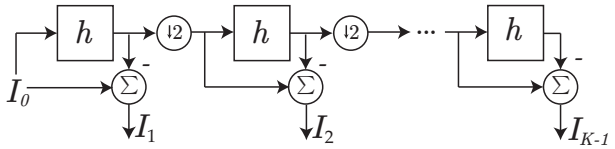

Figure 5.6: Laplacian pyramid of K decomposition levels. $\downarrow 2$ represents the downsampling operation

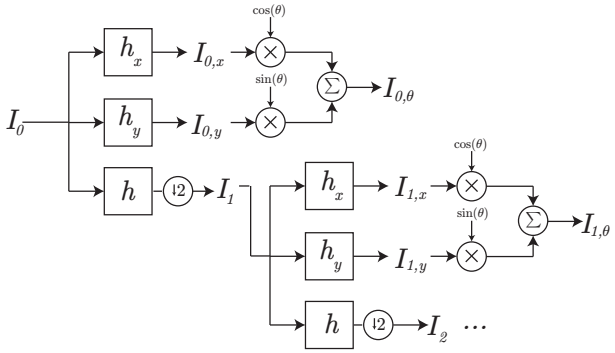

Figure 5.7: Steerable pyramid of 1 decomposition levels using θ as direction.

Steerable pyramid

The steerable pyramid allows to analyze texture at different orientation angles. Note that the steerable pyramid is a direct extension of the Laplacian pyramid in which image details are characterized at different image scales and orientations. Particularly, the steerable pyramid divides an image into a collection of levels localized in both scale and orientation (Freeman and Adelson, 1991; Simoncelli and Freeman, 1995). Thus, an image is decomposed in several scales and each scale is decomposed in a set of directions, which makes this technique suitable for several tasks such as texture analysis/synthesis (Heeger and Bergen, 1995), image enhancement (Wu et al., 1998), image retrieval (Areepongsa et al., 2000), image segmentation (Benjelil et al., 2009), face recognition (Aroussi et al., 2009).

Figure 5.7 shows the block diagram of the steerable pyramid using 1 level in θ direction, where, h, h_x and h_y are a Gaussian filter and its derivatives in x and y direction, respectively. In general, this pyramid is obtained by using Gaussian filters, downsampled images and first derivatives of those images. Thus, an image at the kth scale in the direction θ is defined as $I_{k,\theta} = \cos(\theta)I_{k,x} + \sin(\theta)I_{k,y}$ where $k = \{0, \ldots, K-1\}$ (Lindeberg and Eklundh, 1992) for K levels. Usually, descriptive statistics are computed on the filtered images to characterize the texture.

Laws filters

Laws filters are considered as one of the first filtering approaches in texture analysis, presented by Laws (Laws, 1980). This filtering approach has been used very often as reference for comparing texture analysis techniques. Also, the Laws filters have been used in several texture analysis tasks such as segmentation (Eckstein, 1996), object recognition (Baik and Pachowicz, 2002), image retrieval (Suzuki et al., 2009) and film colorization (Lavvafi et al., 2010). In the following paragraphs a procedure to obtain features using Laws filters is presented. First we introduce the set of filters proposed by Laws. These filters can be constructed using the following one-dimensional kernels:

$$\begin{array}{lcl} L5 & = & [\ \ 1 \ \ \ 4 \ \ 6 \ \ \ 4 \ \ \ 1\ ] \\ E5 & = & [\ -1 \ \ -2 \ \ 0 \ \ \ 2 \ \ \ 1\ ] \\ S5 & = & [\ -1 \ \ \ 0 \ \ 2 \ \ \ 0 \ \ -1\ ] \\ W5 & = & [\ -1 \ \ \ 2 \ \ 0 \ \ -2 \ \ \ 1\ ] \\ R5 & = & [\ \ 1 \ \ -4 \ \ 6 \ \ -4 \ \ \ 1\ ] \end{array} . \tag{5.7}$$

These kernels stand for Level, Edge, Spot, Wave, and Ripple, respectively. From those one-dimensional kernels, 25 different two-dimensional filters are generated, e.g., the $L5E5$ filter is obtained by multiplying a vertical kernel $L5$ and a horizontal kernel $E5$. The filters are used to decompose the image into sub-bands. Those sub-bands are the result of the convolution of the original image and the 25 set of filters. Here, the set of 25 images (one per sub-band) are termed $I_{k,l}$ for $k = \{1, \ldots, 5\}$ and $l = \{1, \ldots, 5\}$. These images are used to obtain the set of Texture Energy Measures (TEM) which are pixel-wise defined as

$$\|I_{k,l}\|(i,j) = \sqrt{\sum_{x=-7}^{7} \sum_{y=-7}^{7} I_{k,l}^2(i+x, j+y)}. \tag{5.8}$$

Then, the features are normalized using the formula $\|I_{k,l}\| = \frac{\|I_{k,l}\|}{\|I_{1,1}\|}$. Descriptive statistics are computed on the resulting TEM to characterize the texture.

Discrete Wavelet transform (DWT)

The wavelet transform has been widely used in texture analysis tasks such as image annotation (Ma and Manjunath, 1995), texture characterization/classification (Mojsilovic et al., 2000; Busch and Boles, 2002; Lam, 2008; Wang and Yong, 2008; Dong and Ma, 2011), defect detection (Han and Shi, 2007), obtaining satisfactory results. Note the close relationship between the steerable pyramid, the Laplacian pyramid and the DWT. This is because they make part of a family of signal processing based techniques where image details are characterized in multiple scales and orientations by using typically intentionally crafted filters (Mallat, 2009). More advanced signal decomposition methods include shearlets (Goossens et al., 2009), contourlets (Do and Vetterli, 2005), ridgelets (Do and Vetterli, 2003), wedgelets (Donoho, 1999).

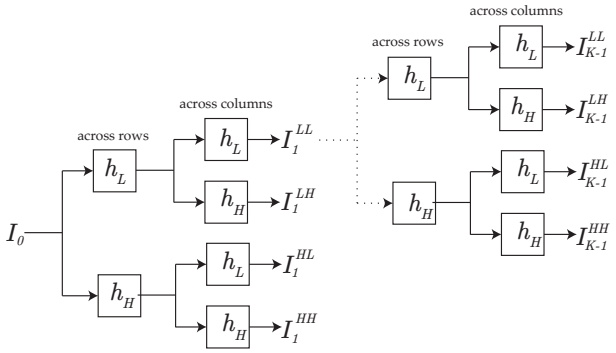

Figure 5.8: Two-dimensional discrete wavelet transform of K decomposition levels.

The wavelet transform has been described as a multi-resolution analysis of a finite energy signal (Mallat, 2009). Particularly, the discrete wavelet transform is a filter bank decomposition of a signal using a low pass and high pass filters (Lam, 2008; Mallat, 2009). Figure 5.8 shows this representation of the DWT of K decomposition levels, where h_L and h_H are low and high pass filters, respectively. Typically, descriptive statistics are computed on every decomposition level to characterize the texture.

Wigner distribution (WD)

The WD is a popular technique to describe the local joint distribution of pixel values in image processing tasks such as image enhancement, image fusion, image segmentation (Homigo and Cristobal, 1998; Cristobal and Hormigo, 1999; Dragoman, 2005; Redondo et al., 2008; Vaidya and Padole, 2008). The discrete approximation is termed the pseudo-Wigner distribution (PWD). The PWD is performed by using two smoothing windows: a spatial averaging window (h_s) and a spatial-frequency averaging window (h_f) (Cristobal and Hormigo, 1999). The PWD of a two-dimensional discrete function is defined as follows

$$\begin{aligned} PWD\{I\}(i,j,u,v) \quad = \quad & \sum_{x=-W/2}^{W/2} \sum_{y=-W/2}^{W/2} h_s(x,y) \\ & \sum_{k=-W/2}^{W/2} \sum_{l=-W/2}^{W/2} h_f(k,l) \\ & I(i+k+x, j+l+y) I^*(i+k-x, j+l-y) \\ & \exp(-2\sqrt{-1}(xu+yv)), \end{aligned}$$

where $W \times W$ is the size of the smoothing window and I^* is the complex conjugate of I. By employing Equation (5.9) in an image I, a set of $W \times W$ images representing the spatial/spatial-frequency distribution of the texture is

Table 5.1: List of texture analysis techniques reviewed in this work.

Approach	Technique
Statistical features	Autocorrelation function (AC)
	Co-occurrence matrix (CM)
	Local Binary Patterns (LBP)
Structural features	Granulometry moments (GM)
Model based features	Autoregressive models (AR)
	Gaussian Markov random field (GMRF)
Signal processing based features	Power spectrum (FFT)
	Eigenfilter (Eig)
	Gabor filters (Gb)
	Laplacian pyramid (LP)
	Laws filters (TEM)
	Steerable pyramid (SP)
	Discrete Wavelet transform (DWT)
	Pseudo-Wigner distribution (PWD)

obtained. The texture is characterized by extracting histograms of the resulting images.

5.2.5 Summary

In Table 5.1, a summary of the considered texture extraction techniques is presented. From the literature review, we conclude that the signal decomposition methods are considered more often than the other approaches. This could be due to the fact that texture is better characterized by means of signal decomposition methods (Xie, 2008). The approaches included in this work were selected by literature review according to their performance or popularity in the image processing field. For instance, Eig, CM, TEM, AR, FFT and LBP are commonly used as benchmark techniques (Ojala et al., 1996; Randen and Husoy, 1999; Singh and Singh, 2002; Pico et al., 2001; Xie, 2008) when comparing texture analysis techniques. DWT and Gb have shown to be accurate for defect detection (Xie, 2008). The other selected techniques were chosen according to their performance in texture classification or retrieval. Particularly, AC is very suitable for fabric inspection (Heilbronner, 1992). LP, SP, GMRF and PWD are very suitable for texture image retrieval and classification (Randen and Husoy, 1999; Areepongsa et al., 2000; Vo et al., 2006). The techniques listed in Table 5.1 are often considered for evaluation of appearance changes in texture in pilling, wrinkling, seam puckering, fuzziness and pile surface assessment (Kang et al., 2005; Hu, 2008; Xiaojun et al., 2009).

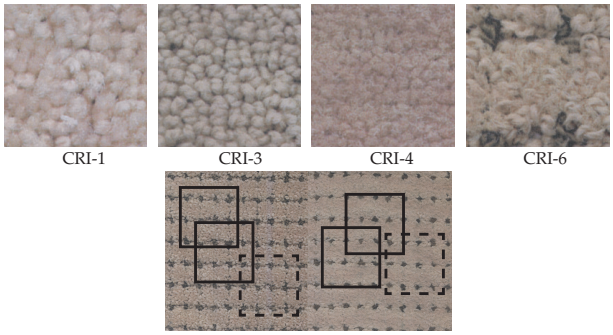

Figure 5.9: (top) Example test textures from the CRI standard. (bottom) Cropping procedure. Example cutouts used as samples.

5.3 Results and discussion

In this Section we describe the test images, the implementation details of the tested methods and the performance comparison between texture descriptors.

5.3.1 Test images

For the interested readers, we provide a description of carpet construction in Appendix A.4. For our experiments, we use three databases, two of them composed of samples of wear with grades in steps of 0.5 from 1.0 to 4.5 and one database which only possesses four samples of wear from 1.0 to 4.0 in steps of 1.0.

The first set of images is composed of scanned printed images, using an office scanner, from the CRI standard photo set. The CRI reference sets include texture types with level loop (CRI-3), cut Saxony (CRI-1 and CRI-4) and tip-sheared loop (CRI-6) piles, see Figure 5.9(top). To keep the relevant characteristics in the scanned images, a resolution of 7.8 pixels per millimeter was used (Orjuela-Vargas et al., 2010). Each printed photograph contains textures corresponding to eight wear degrees from 1.0 to 4.5 in steps of 0.5. Also, in the same image, the original texture of the floor covering is included. Each printed photograph was digitized in a 2300×1100 pixels image (Orjuela-Vargas et al., 2010).

We compute the texture features in cutouts. The cropping procedure is performed by extracting random cutouts from either the part of the original or the part with the appearance change with the purpose of having the reference and test samples, see Figure 5.9(bottom). For this subset we extract 20 cutouts from both reference and test surfaces, we term each pair of cutouts as sample. Therefore, our CRI dataset is composed of 20 samples $\times$ 8 wear labels per reference set $\times$ 4 CRI reference sets, adding up to 640 texture samples (reference plus test surface) in total.

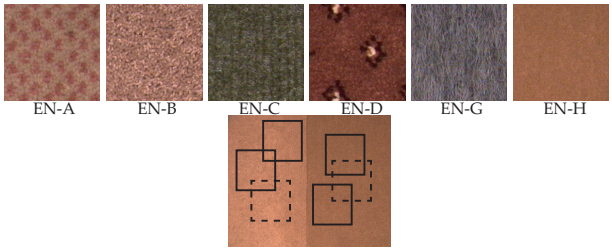

Figure 5.10: (top) Example test textures from the `EN1471` standard. (bottom) Cropping procedure. Example cutouts used as samples.

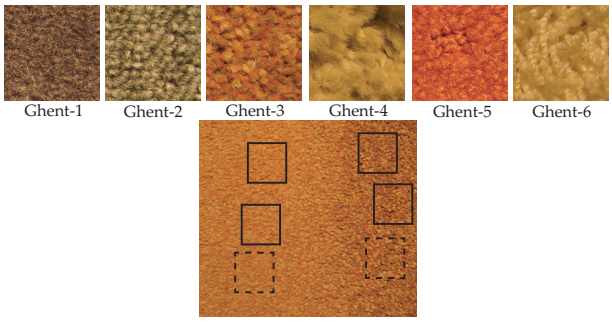

Figure 5.11: (top) Example test textures from the Ghent University Textile Department. (bottom) Cropping procedure. Example cutouts used as samples.

The second set of images is composed of photographs of physical samples from level loop (EN-A), cut Saxony (EN-B, EN-C and EN-D), cut/frisé (EN-G) and woven velours (EN-H) piles from the `EN1471` standard, see Figure 5.10(top). In this set, wear degrees were photographed at $30cm$ with a progressive `3CCD Sony` camera model `DXC-9100 P` using a `Sony` macro lens model `VCL-707BXM`. The database includes photographs of size of 576×720 pixels corresponding to 14.5×18 cm^2. This offers a resolution of 4 pixels per millimeter. Each reference contains photographs corresponding to four wear degrees from 1.0 to 4.0 in steps of 1.0. Also, in the same photograph, the original texture of the floor covering is included (Orjuela-Vargas et al., 2010).

Similarly to the CRI set, the cropping procedure is performed by extracting random cutouts from either the part of the original or the part with the appearance change, see Figure 5.10(bottom). For this subset we extract 20 cutouts from both reference and test surfaces. Therefore, our EN dataset is composed of 20 samples $\times$ 4 wear labels per reference set $\times$ 6 EN reference sets, adding up 480 texture samples (reference plus test surface) in total.

In the third database, the wear degrees were assessed by three inspectors of the textile Department of Ghent University, in collaboration with the textile

floor covering company LANO. The set is composed of one cut pile velours (Ghent-1), one level loop pile (Ghent-2), one cut pile saxony (Ghent-3), two cut pile shag (Ghent-4 and Ghent-6) and one cut/loop pile (Ghent-5), see Figure 5.11(top). Each physical sample corresponds to the eight wear degrees from 1.0 to 4.5 in steps of 0.5. Also, in the same sample, the original texture of the floor covering is preserved. The physical samples were photographed at $30cm$ with a progressive `3CCD Sony` camera model `DXC-9100 P` using a `Sony` macro lens model `VCL-707BXM`. The database includes photographs of size of 720×576 pixels corresponding to $18 \times 14.5\ cm^2$. This offers a resolution of four pixels per millimeter (Orjuela-Vargas, 2012).

The cropping procedure is done similarly to the CRI and the EN sets, see Figure 5.11(bottom). For this subset we extract 20 cutouts from both reference and test surfaces. Therefore, our Ghent dataset is composed of 20 samples $\times$ 8 wear labels per reference set $\times$ 6 Ghent reference sets, adding up 960 texture samples (reference plus test surface) in total.

Note that we use the cropping procedure with the purpose of increasing the number of samples in the experiment (we analyze many cropped regions instead of one large image) and providing more general conclusions.

5.3.2 Implementation details

In this Section we explore the parameter selection for the techniques described in Section 5.2.

First we select an appropriated measure to compute the dissimilarity or difference between the extracted features of a used textile sample and a reference (new textile). In this work, we use the Euclidean distance and the symmetrized adaptation of the Kullback-Leibler divergence (KLD). On the one hand, the Euclidean distance was selected to measure differences in the following techniques: AC, CM, AR, GMRF, GM, FFT, Eig, LP, SP, TEM and DWT. The Euclidean distance was selected for measuring differences in those techniques because it is probably the most common chosen type of distance due to its simplicity. Also, the Euclidean distance can be used to model numerous natural facts of the human-scale world and most of the powerful image recognition techniques make use of it (Gan et al., 2007). On the other hand, we use the symmetrized adaptation of the KLD in the following techniques: LBP, Gb and PWD. This divergence is used to measure differences in the above techniques because those generate a histogram and the symmetrized adaptation of the KLD has proved to be very accurate in measuring differences between histograms. Also, it has proved to be very suitable for texture analysis applications (Dong and Ma, 2011).

In the AC technique we use the parametric model described by (Petrou and Sevilla, 2006). In such a model, texture descriptors are represented by the coefficients of a two dimensional second order polynomial. The texture difference/dissimilarity is computed as the Euclidean distance between the obtained coefficients of the textures under analysis.

In the CM, we use the displacements $(x, y) \in \{(0,1),(-1,1),(-1,0),(-1,-1)\}$ suggested in (Randen and Husoy, 1999). For each displacement in the set, a matrix of frequencies is obtained. The four obtained matrices are averaged to obtain a single co-occurrence matrix. The average is used because it has shown to be more accurate than using the individual matrices (Popescu et al., 2007). The following measures are computed on the average co-occurrence matrix to extract texture features: energy, entropy, contrast, homogeneity and correlation (Tuceryan and Jain, 1998; Randen and Husoy, 1999). The texture difference is computed as the Euclidean distance between the obtained features.

In the LBP technique, the texture units are grouped into a histogram to extract texture descriptors (Maenpaa, 2003). The texture differences are computed by using the symmetrized adaptation of the KLD. Note that in this texture analysis method, there are 2 parameters that influence the output of the LBP codes. The first one is the radius of the selected neighborhood and the second is the threshold ϵ. These parameters are explored later in Section 5.3.3.

The estimated parameters of the AR model are used as texture descriptors. Here the radius of the selected neighborhood is also a free parameter and it is explored later in Section 5.3.3. Similarly, the parameters estimated in the GMRF model are used as texture features. The texture difference is computed as the Euclidean distance between the obtained features in these two techniques.

We compute the following descriptive statistics known as the *granulometry* and *anti-granulometry* moments (GM and anti-GM) in the set of images obtained by *granulometry* and *anti-granulometry* technique: the average, variance, skewness and kurtosis (Aptoula and Lefevre, 2011). In this technique the number of successive morphological operations, termed λ, is kept as free parameter and it will be explored later in Section 5.3.3. The texture difference is computed as the Euclidean distance between the obtained *granulometry* and *anti-granulometry* moments.

In the *power spectrum* technique, the set of wedge and ring filters proposed by (Weszka et al., 1976) is used. Particularly, the following set of parameters $r_1 = \{2,4,8,16,32,64\}$, $r_2 = \{4,8,16,32,64,128\}$, $\theta_1 = \{112.5, 67.5, 22.5, 157.5\}^\circ$ and $\theta_2 = \{247.5, 247.5, 202.5, 292.5\}^\circ$. The energy of the image in each frequency band generated by the filters is computed as texture features. The texture difference is computed as the Euclidean distance between the obtained features.

In the Eig method, the mean and standard deviation of the resulting subbands are computed as feature vectors. Afterwards, the ED is computed as texture difference. If a circular neighborhood is used, the radius of the selected neighborhood is also a free parameter for this technique and it is explored as well later in Section 5.3.3.

In the case of Gb technique, we use the configuration proposed by (Manjunath and Ma, 1996) with the purpose of reducing the redundancy presented in the filter bank decomposition. Particularly, the following parameters are used:

$U_h = 0.7$, $U_l = 0.005$, $K = 6$ and $S = 4$. Afterwards, the histogram of each sub-band is used as texture descriptor. The symmetrized adaptation of the KLD is computed as dissimilarity measure in this technique.

In the LP and SP techniques, the feature vectors are the mean and standard deviation of the resulting image decomposition. We use the following set of angles for the steerable pyramid: $\{0°, 45°, 90°, 135°\}$. Here the number of scales remains as free parameter to be explored later in Section 5.3.3. In TEM technique, the mean and standard deviation of the resulting filtered images are computed as feature vectors. Also for the DWT technique, the mean and standard deviation of the resulting decomposition are computed as texture descriptors. Here the wavelet function and the number of scales remain as free parameters to be explored later in Section 5.3.3. The texture difference is computed as the Euclidean distance between the obtained features in these four techniques.

For the PWD method, the histogram of each sub-band is used as texture descriptor. The texture differences are computed using the symmetrized adaptation of the KLD between the histograms of the resulting PWD.

5.3.3 Impact of the parameters

In this Section we study the impact of the free parameters of the tested techniques on the performance. That is, we compute independently the correlation between the texture feature differences and wear labels for each technique and selected parameter values. Particularly, we study the following parameters:

- In the AR model and Eig technique we study the radius $r \in \{1, 1.5, 2, 3\}$.
- In the LBP technique we study the radius $r \in \{1, 1.5, 2, 3\}$ and the threshold $\epsilon \in [0, 0.05]$.
- In the LP and SP techniques, we study the number of scales $K \in \{1, 2, 3, 4, 5\}$.
- In GM technique, we study the number of successive morphological operations $\lambda \in \{2, 3, \ldots, 27, 28\}$.
- In the DWT technique, we study the wavelet filters, particularly the set of Daubechies wavelet filters, and the number of scales $K \in \{1, 2, 3, 4, 5\}$. We use the notation dbN to identify the Daubechies filter with N filter coefficients (N$\in \{1, 2, \ldots, 19, 20\}$).

Figure 5.12 shows the performance (PCC, SROCC and CCD, see Section 2.3) of the AR model in function of the radius r. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). The correlation is computed independently for each reference set. The Figure shows that the AR model achieves the highest correlation with the textile specialists' assessment for the radius $r = 1.5$. However, the AR model exhibits a weak correlation (correlation

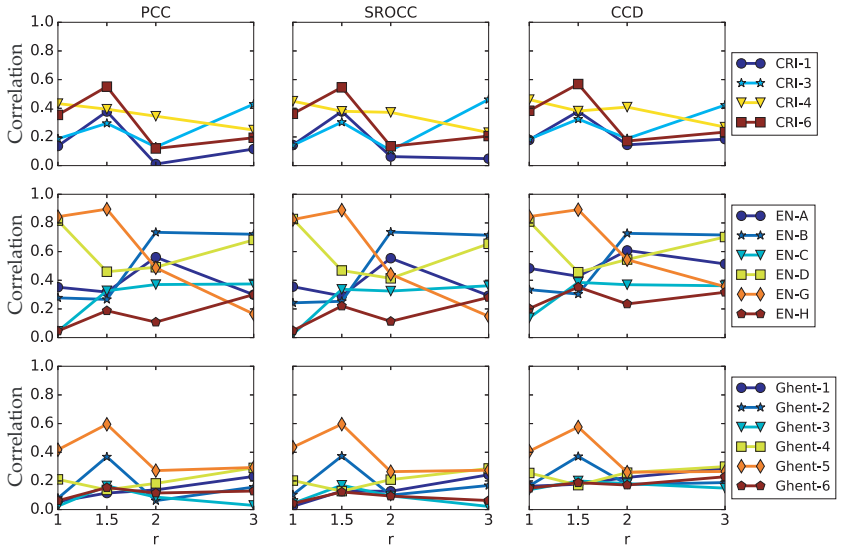

Figure 5.12: Performance of the AR model in function of the radius r. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD.

between texture feature differences and wear labels lower than 0.5) in 14 out of the 16 evaluated reference sets. Thus, the AR model does not provide good texture features to evaluate appearance changes in texture.

Figure 5.13 shows the performance (correlation between the texture feature differences and the textile specialists' assessment) of the Eig technique in function of the radius r. The correlation is computed independently for each reference set. The Figure shows that the Eig technique does not exhibit significant differences in terms of performance between the different r values. However, the radius $r = 1.5$ is a good choice for the 16 reference sets. The Eig technique exhibits a strong correlation (correlation between texture feature differences and wear labels higher than 0.7) in 10 out of the 16 evaluated reference sets (CRI-1, CRI-3, EN-A, EN-B, EN-C, EN-D, EN-G, EN-H, Ghent-3 and Ghent-5). Particularly, the Eig technique performs well in the cut-pile and loop-pile types of floor coverings for the three datasets.

Figure 5.14 shows the performance (correlation between the texture feature differences and the textile specialists' assessment) of the GM technique in function of the λ values. The correlation is computed independently for each reference set. The method achieves the highest correlation (SROCC) for $\lambda = 14$ in the EN set as well as in the Ghent set and for $\lambda = 26$ in the CRI set. However, the GM technique exhibits a weak correlation (correlation between texture feature differences and wear labels lower than 0.5) in 12 out of the 16 evaluated reference sets. Thus, this technique is not a good texture

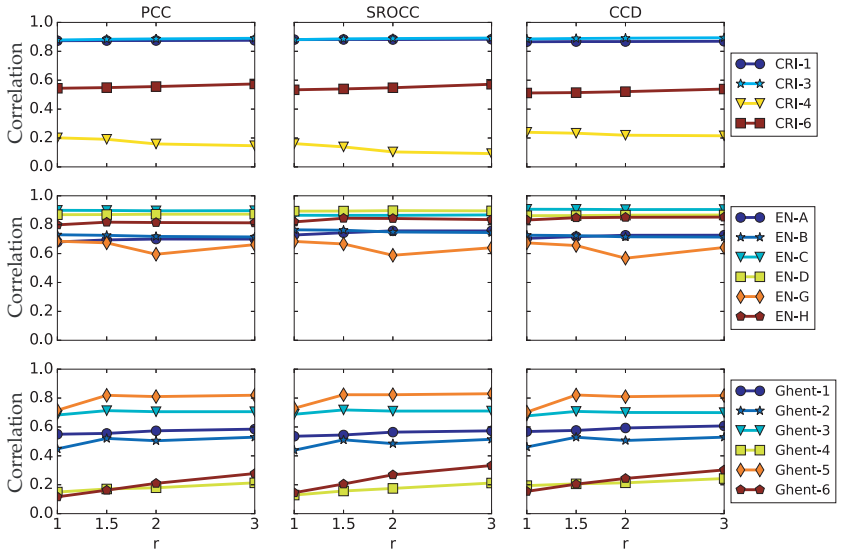

Figure 5.13: Performance of the Eig technique in function of the radius r. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD.

feature to evaluate appearance changes in texture. This could be because from the structural point of view, texture is characterized by primitives and spatial arrangement of those primitives (Tuceryan and Jain, 1998; Xie, 2008). Typically, any variation from the textural primitives of the reference (new textile) is considered as a different texture but the magnitude of the variation does not indicate how big or small is the difference of the used textile compared to the new textile. Therefore, this kind of algorithms is limited in power to discriminate between very similar textures, e.g., evaluation of appearance changes in textile floor coverings due to degradation. This aspect makes techniques based on structural primitives impractical for evaluating appearance changes in texture (Tuceryan and Jain, 1998).

Figures 5.15 and 5.16 show the performance (correlation between the texture feature differences and the textile specialists' assessment) of the LP and SP techniques in function of the scale K. The correlation is computed independently for each reference set. Both methods achieve the highest performance at the highest decomposition levels ($K > 4$). The LP and SP techniques exhibit a strong correlation (correlation between texture feature differences and wear labels higher than 0.8) in the EN subset (EN-A, EN-B, EN-C, EN-D, EN-G and EN-H). The correlation between the texture differences and the wear labels is at most moderate (lower than 0.7) in the CRI and Ghent databases.

Figures 5.17, 5.18 and 5.19 show the performance (correlation map) of the

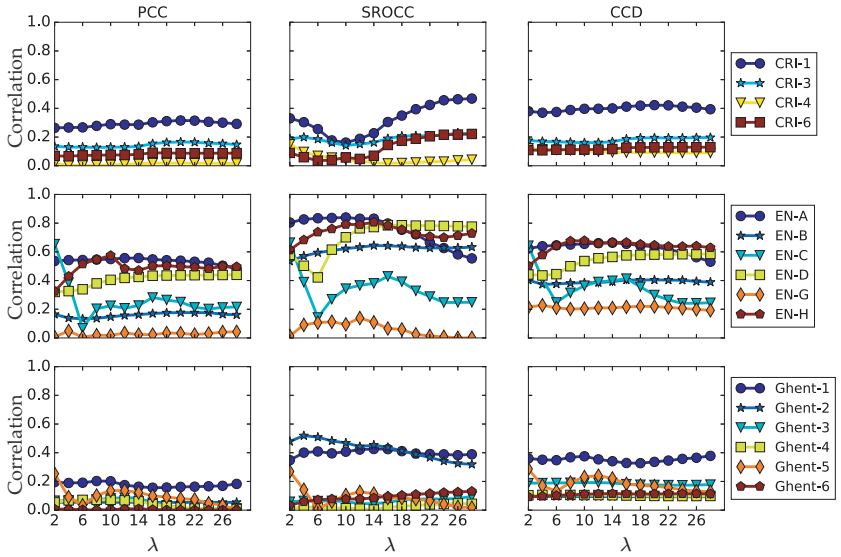

Figure 5.14: Performance of the GM technique in function of the λ values. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD.

LBP technique in function of the radius r (y-axis) and threshold ϵ (x-axis) appraised on CRI, EN and Ghent subsets, respectively. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). The correlation is computed independently for each reference set. The color bar represents the strength of the correlation in the color map from 0 to 1. Note that the best performance (highest correlation between texture feature differences and wear labels) is achieved for a threshold $\epsilon = 0.01$ and the radius does not have much impact for this threshold level, i.e., there are not significant changes in terms of correlation for this threshold level. However, for bigger ϵ values, the LBP technique achieves a higher correlation for $r = 3$. The LBP technique exhibits a strong correlation (correlation between texture feature differences and wear labels higher than 0.8) in 7 out of the 16 evaluated reference sets (CRI-1, CRI-3, CRI-6, EN-B, EN-C, EN-D and Ghent-1). Particularly, the method shows a good performance in the cut-pile types of floor coverings for all three datasets.

Figures 5.20, 5.21 and 5.22 show the performance (correlation map) of the DWT technique in function of the scale K (y-axis) and number of filter coefficients N (x-axis) appraised on CRI, EN and Ghent subsets, respectively. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). The correlation is computed independently for each reference set. The color bar represents the

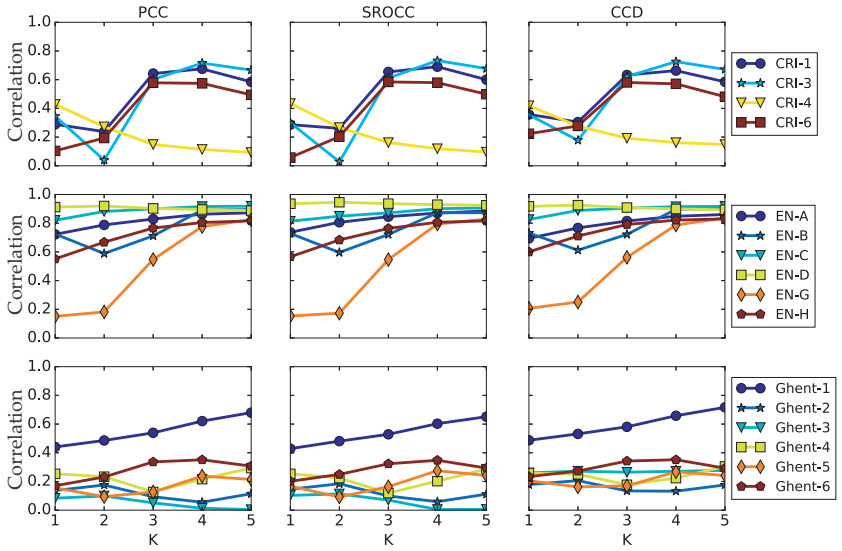

Figure 5.15: Performance of the LP technique in function of the scale K. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD.

strength of the correlation in the color map from 0 to 1. The best performing (highest correlation with wear labels) wavelet is the Haar wavelet (Daubechies filter with 2 coefficients [db1]). The DWT technique achieves the highest correlation between texture feature differences and wear labels for $K = 4$ decomposition levels. Note that higher decomposition levels result in lower performance. The DWT technique exhibits strong correlation (correlation between texture feature differences and wear labels higher than 0.8) in 12 out of the 16 evaluated reference sets (CRI-1, CRI-3, EN-A, EN-B, EN-C, EN-D, EN-G, EN-H, Ghent-1, Ghent-2, Ghent-3 and Ghent-5). That is, the method shows a good performance in the cut-pile and loop-pile types of the tested floor coverings.

Based on the previous results we select the following set of parameters:

- In the AR model and Eig techniques, we selected $r = 1.5$ as the radius of the neighborhood.
- In the LBP technique, we selected the radius $r = 1.5$ and the threshold $\epsilon = 0.01$.
- In the LP and SP techniques, we selected $K = 4$ as the number of scales.
- In GM technique, we select $\lambda = 20$ as the number of successive morphological operations.

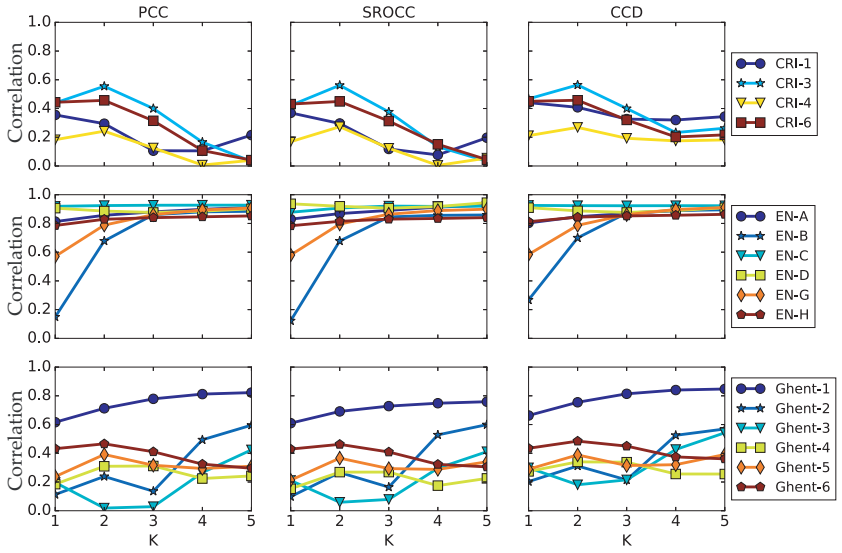

Figure 5.16: Performance of the SP technique in function of the scale K. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD.

- In the DWT technique, we selected the Haar wavelet with $K = 4$ decomposition levels.

The rest of the selected parameters were discussed in Section 5.3.2

5.3.4 Performance comparison

In the following paragraphs we compare the performance between the studied methods using the previously selected parameters.

Figure 5.23 shows the performance of the considered texture analysis techniques appraised on (a) CRI, (b) EN and (c) Ghent databases. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). The correlation is computed for all reference sets, i.e., we have included all the data samples for computing the correlation between wear labels and texture feature differences. In the EN database, DWT, LP and SP techniques provide good descriptors in assessing changes of texture in textile floor coverings, when a labeling error of 1 between consecutive wear labels is allowed, displaying a strong correlation (correlation between texture feature differences and wear labels higher than 0.7). However, when the allowed labeling error between consecutive wear labels is 0.5, CRI and Ghent databases, the tested texture analysis techniques display a weak correlation (correlation between texture feature differences and wear labels lower than

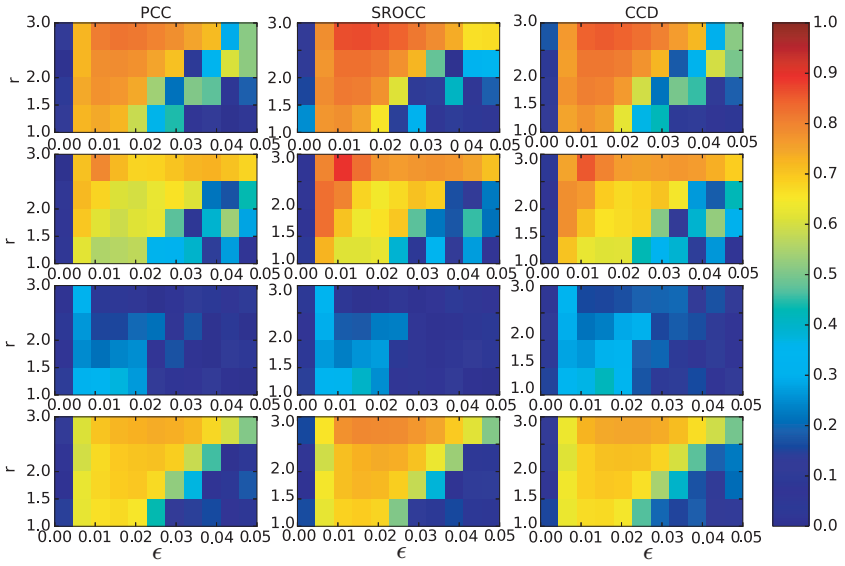

Figure 5.17: Performance of the LBP technique in function of the radius r and threshold ϵ appraised on CRI subset. From top to bottom: CRI-1, CRI-3, CRI-4 and CRI-6. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD. The color bar represents the strength of the correlation from 0 to 1.

0.5) when evaluating all the reference sets at once.

Since the texture patterns for the floor coverings are standardized in the textile industry (see Appendix A.4), it is also interesting to study the individual correlation for each standard reference set (Orjuela-Vargas, 2012). Therefore, box plot analysis is very useful for identifying how well the measures perform if the image content remains the same (in this case the reference set).

Figure 5.24 shows the box plot for the considered texture analysis techniques per reference set appraised on (a) CRI, (b) EN and (c) Ghent databases. The box plot is a graphical representation of the 4, 6 and 6 PCCs, SROCCs and CCDs of CRI, EN and Ghent databases computed between the texture feature differences and the wear labels for each reference set. The correlation is computed independently for each reference set. That is, for each individual reference set, we assess the correlation between the texture feature differences and the textile specialists' assessment over all wear labels. Thus, it shows the variability of the agreement between the tested texture analysis techniques and the wear labels under different reference set (texture content), i.e., it is an indication of how well the technique performs for the different reference sets (texture pattern).

We select from the set of texture analysis techniques, the techniques with the

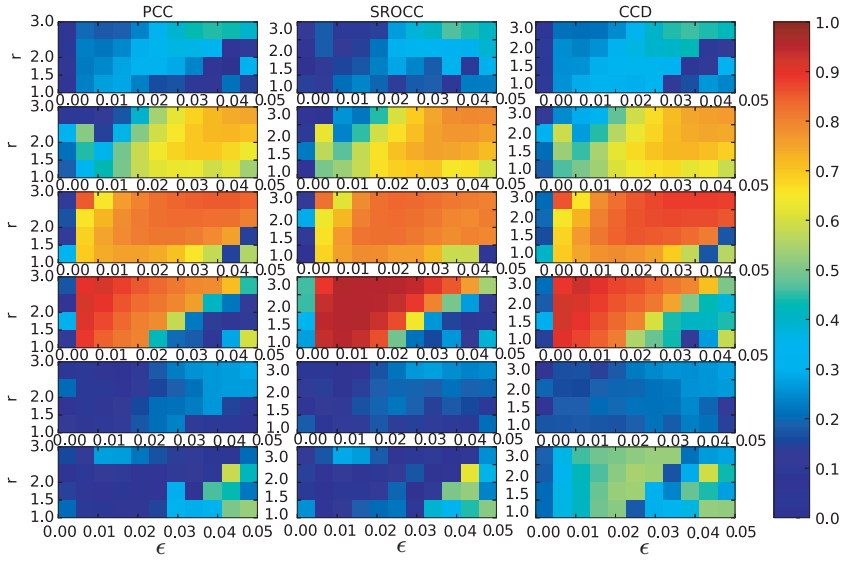

Figure 5.18: Performance of the LBP technique in function of the radius r and threshold ϵ appraised on EN subset. From top to bottom: EN-A, EN-B, EN-C, EN-D, EN-G and EN-H. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD. The color bar represents the strength of the correlation from 0 to 1.

best performance by means of statistical analysis. Specifically, we use multiple statistical comparisons as discussed in (Garcia et al., 2010b) (see Section 2.3). The objective of this test is to determine if we may conclude from the correlation values of each reference set that there are differences among the tested texture analysis techniques for the three databases. From the multiple statistical comparisons we found that the best performing functions are (p-values <0.05):

- for EN: DWT is statistically significant better than: AC, AR, CM, FFT, GAM, GMRF, GB, LBP and PWD;
- for CRI: DWT is statistically significant better than: AR, FFT, GAM, LP, PWD and SP;
- for Ghent: DWT is statistically significant better than: AR, CM, GAM, GMRF, LBP, LP, PWD and TEM.

Note once more that the highest overall performance of DWT, LP and SP techniques is achieved on the EN database (see Figure 5.24(b)). The box plots are characterized by strong median correlation, short boxes (box length <0.05 in the correlation scale) and short whiskers. This suggests once more that

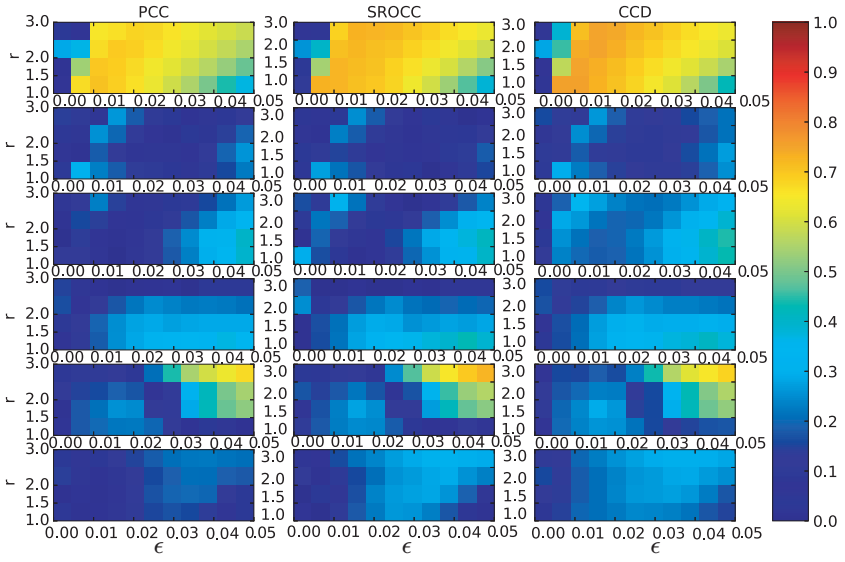

Figure 5.19: Performance of the LBP technique in function of the radius r and threshold ϵ appraised on Ghent subset. From top to bottom: Ghent-1, Ghent-2, Ghent-3, Ghent-4, Ghent-5 and Ghent-6. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD. The color bar represents the strength of the correlation from 0 to 1.

DWT, LP and SP techniques are good candidates for assessing appearance changes in textile floor coverings when labeling errors of 1 are allowed. That is, these three techniques can be used for making a distinction between carpets suitable for domestic end use (wear label between 2-3) and carpets suitable for commercial end use (wear label between 3-5) (ISO-10361:2000, 2005). However, after detecting the end use, domestic or commercial, these three techniques cannot further differentiate between the subcategories: light use, medium use or intensive use which in general needs wear samples in steps of 0.5 ([2, 2.5, 3] for domestic use and [3, 3.5, 4] for commercial use).

The box plots of CRI and Ghent databases (Figure 5.24(a) and (c)) are characterized by large size boxes (box length greater than 0.2 in the correlation scale) and large whiskers. This is an indication that the texture analysis techniques do not perform well for all the different reference sets over the whole range of wear levels (wear samples in steps of 0.5). That is, the tested techniques perform well only in some of the reference sets of CRI and Ghent databases. Particularly, the DWT has shown to be a good technique for assessing changes of texture in the following reference sets: CRI-1, CRI-3, EN-A, EN-B, EN-C, EN-D, EN-G, EN-H, Ghent-1, Ghent-2, Ghent-3 and Ghent-5. That is, the DWT shows a good performance in the cut-pile and loop-pile

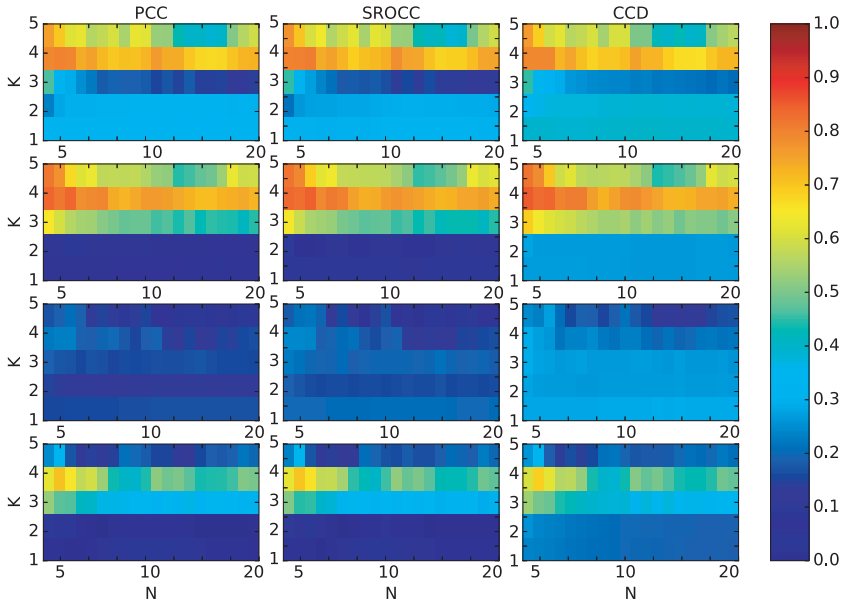

Figure 5.20: Performance of the DWT technique in function of the scale K and number of filter coefficients N appraised on CRI subset. From top to bottom CRI-1, CRI-3, CRI-4 and CRI-6. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD. The color bar represents the strength of the correlation from 0 to 1.

types of the tested floor coverings and it can be used for making a distinction between carpets suitable for domestic and commercial end use as well as for differentiating between the subcategories: light use, medium use or intensive use.

Note that overall the DWT technique is the best performing texture analysis technique, i.e., it performs equally or significant better than the other tested techniques. In general, DWT, Eig, FFT, Gb and LBP are the best performing texture analysis techniques with strong correlations for 10 or more reference sets (CRI-1, CRI-3, EN-A, EN-B, EN-C, EN-D, EN-G, EN-H, Ghent-3 and Ghent-5). That is, these texture analysis techniques show good performance in the cut-pile and loop-pile types of the tested floor covering databases.

The results show that the signal processing methods are the best performing for assessing appearance changes in texture. These methods perform well with a strong correlation (correlation between the texture feature differences and the textile specialists' assessment higher than 0.8) in cut (excluding cut pile shag carpets [Ghent 4 and 6]) and loop pile surface constructions. This result is an important step toward the development of an automatic grading system for cut and loop pile surface constructions, complying with international

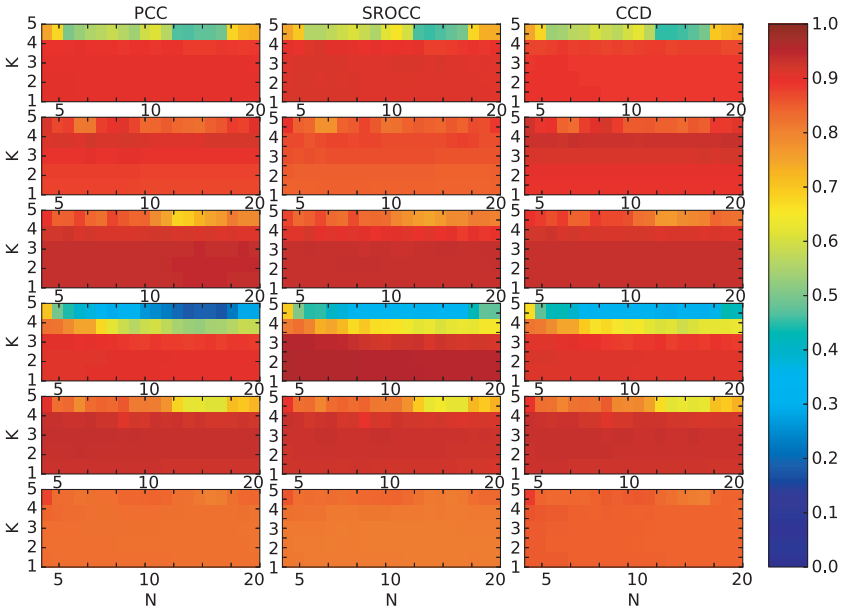

Figure 5.21: Performance of the DWT technique in function of the scale K and number of filter coefficients N appraised on EN subset. From top to bottom EN-A, EN-B, EN-C, EN-D, EN-G and EN-H. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD. The color bar represents the strength of the correlation from 0 to 1.

standardizations, for evaluating appearance changes in textile floor coverings.

Note that the signal processing based methods only show a weak correlation (correlation between the texture feature differences and the wear labels lower than 0.5) in the floor coverings with patterns created using combination of cut and loop piles (CRI-6 and Ghent-5) and the shag pile surface constructions (long piles [about 2.5 cm or higher]). On the one hand, carpets with cut pile shag surface construction exhibit the changes due to wear in a characteristic termed hairiness (Quinones-Lara et al., 2011), which is more related to the edges than the texture. That is, carpets with cut pile shag surface construction cannot be automatically assessed by means of texture image analysis. On the other hand, carpets CRI-6 and Ghent-5 have patterns that are difficult to characterize by means of texture analysis. Therefore, a technique to analyze these patterns is necessary for improving the results in these reference sets, e.g., a measure to compare geometric features of the pattern shapes. Another way of improving the results presented in this Chapter, particularly for those methods using the ED, is by exploring different dissimilarity measures or by learning the distance by using distance learning techniques.

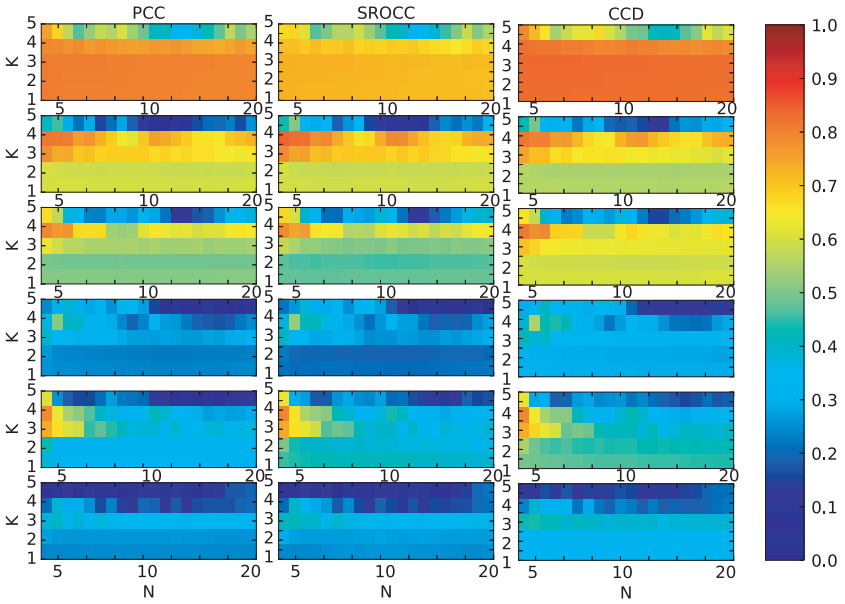

Figure 5.22: Performance of the DWT technique in function of the scale K and number of filter coefficients N appraised on EN subset. From top to bottom Ghent-1, Ghent-2, Ghent-3, Ghent-4, Ghent-5 and Ghent-6. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels). From left to right: PCC, SROCC and CCD. The color bar represents the strength of the correlation from 0 to 1.

5.4 Conclusions

This Chapter has reviewed and evaluated features in the assessment of appearance changes in texture. Particularly, we investigated the problem of appearance change in textile floor coverings due to degradation. We evaluated fourteen texture descriptors for characterizing changes in texture due to wear. We included descriptors based on statistics, filtering, structural and models. Additionally, we have studied the impact on the performance of the free parameters on the tested techniques. To stimulate further experimentation, we made all the tested methods freely available as a plugin on the iFAS software tool. The wear degree was quantified by using descriptor differences between a reference sample and a degraded specimen. The results showed that the DWT, Eig, FFT and Gb techniques provide good descriptors in measuring appearance changes in floor coverings. Particularly, the signal processing methods are the best performing with a strong correlation (correlation between the texture feature differences and the textile specialists' assessment higher than 0.8) in cut and loop pile surface constructions. Therefore, we believe that future work in the evaluation of appearance changes in texture should be developed by using

signal processing methods. The results also showed that the tested texture analysis techniques perform poorly in textile floor coverings with (shag) long pile construction.

The study of other descriptors to texture for improving the process of measuring appearance retention in floor coverings remains a future work. For instance, it is necessary to explore descriptors as hairiness, pilling, change of pattern definition, change in color, among others for improving the results presented in this Chapter. Additionally, since it is common to combine signal processing approaches with other methods, e.g., model based approaches, the study of the combination of different techniques remains as future work with the purpose of potentially improving the process of measuring appearance retention of textiles. For instance, it could be beneficial to model wavelet sub-bands by using MRF (Fan and Xia, 2003) or autoregressive model (Yazdani and Andani, 2017) in order to improve the results presented in this thesis. Also, it is possible to use machine learning techniques to combine multiple texture feature differences to estimate the degradation of the textile under analysis, e.g., linear regression (Ortiz-Jaramillo et al., 2014b), neural networks (Song et al., 2016), support vector machines (Wahba et al., 2017), among others.

In addition, since the methodology is quite generic, it can be applied in any application which requires a comparison between global textures features (evaluation of appearance changes in texture). For instance, wrinkling assessment, pilling assessment, seam puckering, fuzziness, among others.

The contributions reported in this Chapter resulted in one international conference proceedings (Ortiz-Jaramillo et al., 2017) and one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2014b).

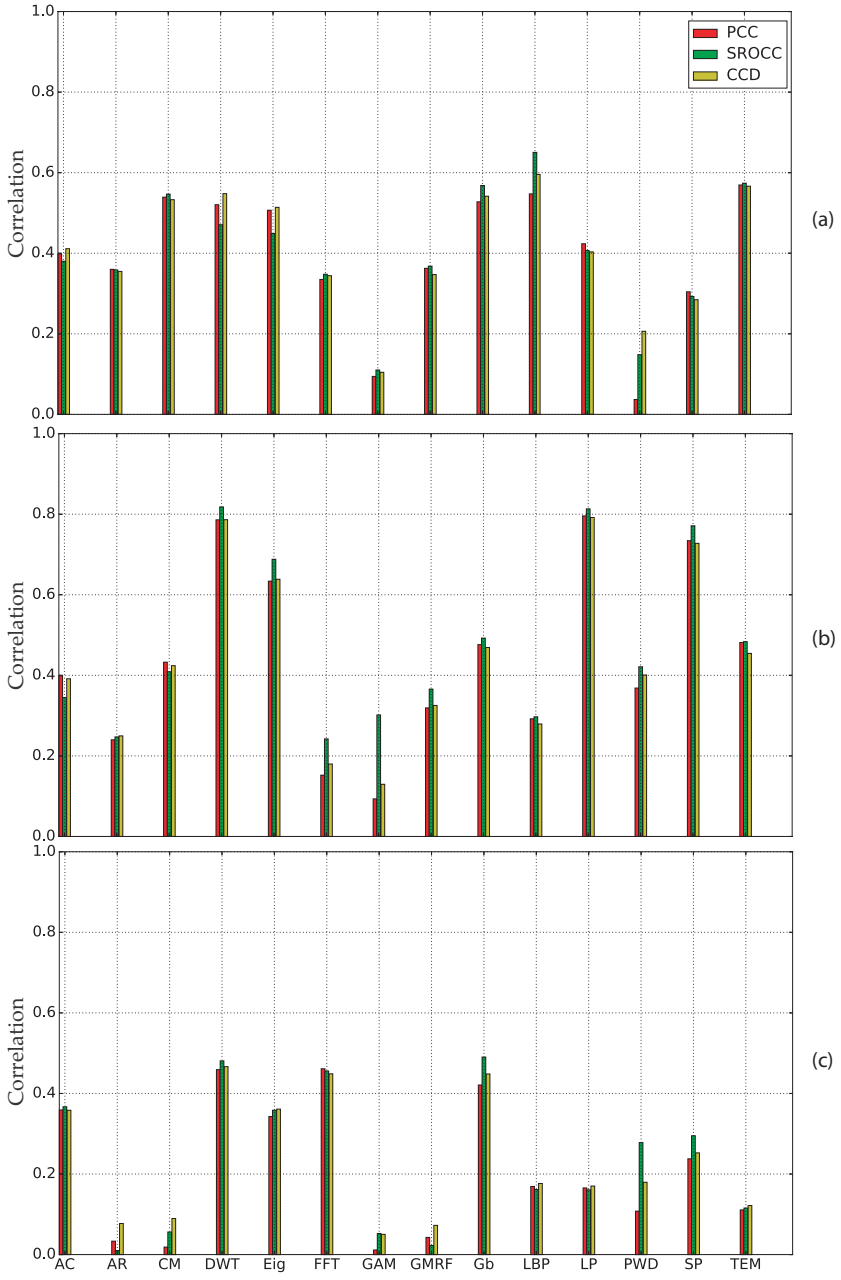

Figure 5.23: Performance of the considered texture analysis techniques appraised on (a) CRI, (b) EN and (c) Ghent databases. The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels).

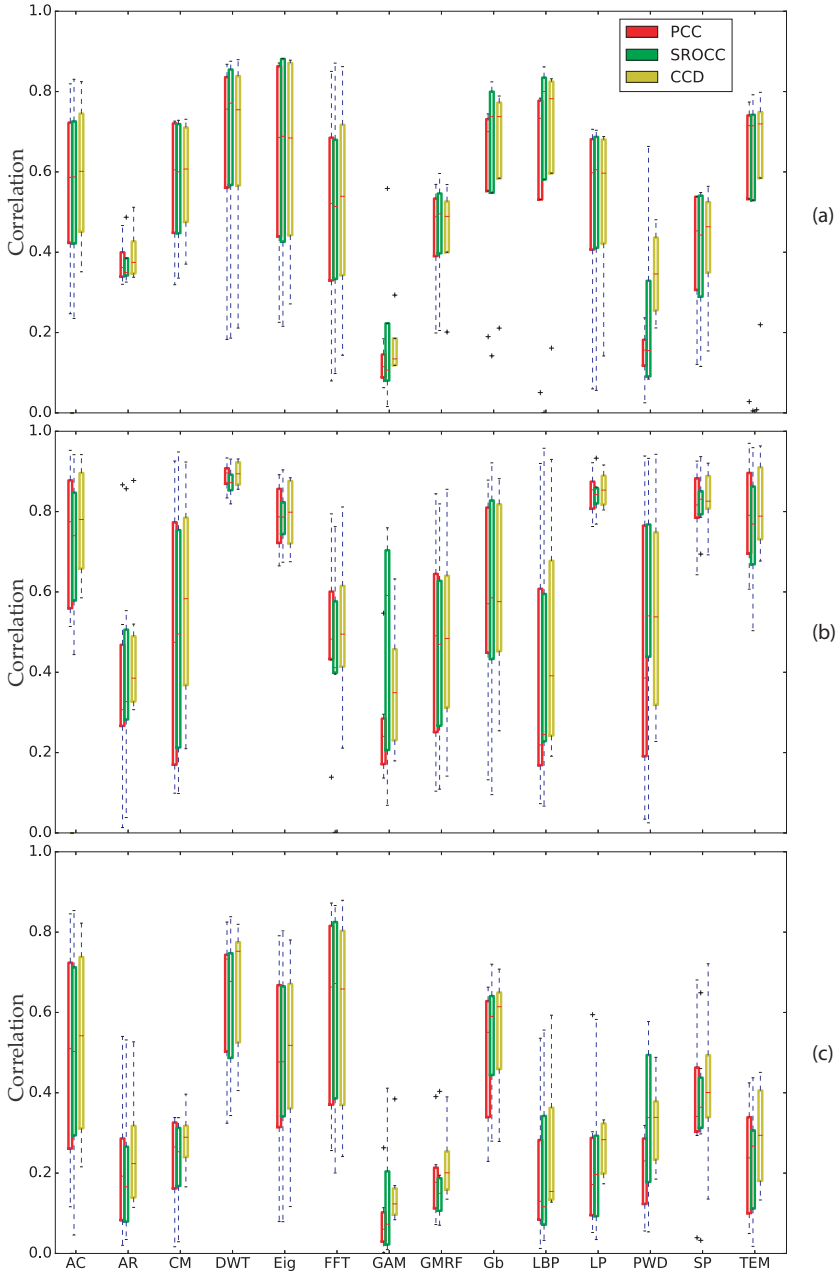

Figure 5.24: Performance of the considered texture analysis techniques appraised on (a) CRI, (b) EN and (c) Ghent databases per reference sets. The box plot was created using the (a) 4, (b) 6 and (c) 6 PCCs, SROCCs and CCDs (one for each reference set). The performance is given in terms of correlation between the texture feature differences and the textile specialists' assessment (wear labels).

6

Evaluation of color differences in natural scene images

6.1 Introduction

Nowadays, fidelity assessment of images in terms of color or simply assessment of color differences (CDs) in images has become an active area in the research of color science and imaging technology due to its wide range of applications such as color correction (Fezza et al., 2014; Ly et al., 2015), color quantization (Brun and Tremeau, 2002), color mapping (Morovic, 2008), color image similarity and retrieval (Mojsilovic et al., 2002). For instance, in multiview imaging, color correction is used to eliminate color inconsistencies between views. In that application, the fidelity assessment of color corrected images relative to the current view image can be used to select the color correction algorithm that produces the smallest perceived color differences. In color mapping and color quantization algorithms, pixel colors are replaced following certain criteria while they ensure a good correspondence in terms of perceived color between the original image and its reproduction. There, CD assessment can be used to find the appropriate quantization step size and/or range of displayable colors to obtain the reproduction with the minimum perceived CD. Another example is color image similarity and retrieval where all images with color composition similar to the query image are retrieved from a database. Thus, the assessment of CDs between images is very important to identify the images with color content similar to that of the query image.

While many CD measures for natural scene color images have been proposed, there has not yet been any rigorous investigation into the performance comparison of the existing measures (Hasler and Susstrunk, 2003; Hardeberg et al., 2008; Rajashekar et al., 2009; Yang et al., 2012; Lee and Rogers, 2014). The CD measures in the state-of-the-art are often tested on databases which: (1) contain multiple distortions in combination with the color-related distor-

tions, (2) include few test image samples, and/or (3) are not publicly available but rather kept private. Additionally, the performance of the CD measures is often reported as average performance over all distortion types of a given database. Overall, to the best of our knowledge, there is little research addressing the problem of reviewing and especially testing CD measures and the existing reports are very limited in test samples and/or CD measures.

In order to address the limitations of the current literature, we take into account various types of CD measures and test those using a public image database which addresses specifically color related image alterations. Specifically, our analysis includes 25 source images which leads to more generalizable results compared to the 6 or 8 source images presented in the other related works (Bando et al., 2005; Hardeberg et al., 2008; Kivinen et al., 2010; Xu et al., 2013). Also, this work includes a list of eighteen CD measures. We made these measures freely available as a plugin on the iFAS software tool (see Appendix B). Firstly, we conduct a brief review in color science for evaluating CDs. Thereafter, we evaluate the eighteen state-of-the-art CD measures and discuss their performances as well as investigate the specific cases where the CD measures fail in order to objectively assess the strengths and weaknesses of the tested measures.

Additionally, we propose a novel method to compute color differences in natural scene color images based on the findings of the review. We base our measure on the fact that humans assess color differences in natural scene color images by comparing sets of connected pixels or small patches. Those patches are typically characterized for being homogeneous or for possessing an unique texture pattern. Therefore, we use image segmentation based on texture to compute the color differences in the resulting segments. Particularly, we use the Local Binary Patterns as texture descriptor because of its simplicity while being one of the most accurate texture analysis algorithms (Maenpaa, 2003). To compute the color differences we use the statistics proposed in (Pinson and Wolf, 2004a) because they are good measures of the change in the color distribution spread and severe color differences. For computing the intensity differences, we use the well-known structural similarity index measure (SSIM) (Zhou et al., 2014). Finally, the overall color difference is computed as the weighted average of the local differences using as weights the ratio between the number of pixels in the patch and the total number of pixels in the image.

We have tested our measure as well as the state-of-the-art measures on three color related distortions (mean shift, change in color saturation and quantization noise) from one image quality assessment database (TID2013 (Ponomarenko et al., 2015)). We found that the proposed measure is able to accurately predict the color differences typically perceived and reported by a human observer. Particularly, our experimental results show that the correlation between the subjective scores and the proposed measure exceeds 80% which is better than the other eighteen CD measures tested in this work. For illustration the best performing state-of-the-art CD measures achieve correlation with humans scores lower than 75%.

This Chapter is organized as follows. In Section 6.2, current approaches dealing with CD assessment in natural scene color images are discussed. The novel methodology is described in Section 6.3. Thereafter, we present and discuss the results obtained in our experimental study in Section 6.4. Finally, we draw conclusions and propose future work in Section 6.5.

6.2 Background

The Commission Internationale de l'Eclairage (CIE) defines color as: "*attribute of visual perception consisting of any combination of chromatic and achromatic content.*" The definition implies that color is an attribute of visual perception, i.e., the study of color is mostly about perception (color appearance) (Johnson and Fairchild, 2002). The study of color appearance seeks to describe the perceptual aspects of human color vision. For instance, the most successful color appearance model (CAM) in the state-of-the-art is the CIELAB (Moroney et al., 2002; Habekost, 2013). Therefore, most of the CD formulas use a certain distance measure in the CIELAB color space (Sharma, 2002). Next to the CIELAB, also other CAMs have been proposed in the state-of-the-art such as $\mathrm{YC_BC_R}$ (ITU, 1995), HSI (Smith, 1978), $\ell\alpha\beta$ (Ruderman et al., 1998), CIELUV (CIE, 1976), OSA-UCS (Huertas et al., 2006). Further information about CAMs can be found in (Sharma, 2002; Johnson and Fairchild, 2002; Mandic et al., 2006; Habekost, 2013).

6.2.1 Color difference measures in images

The most well-known and widely used CD measures for natural scene color images are listed in Table 6.1 and described in the following paragraphs.

Just noticeable CD measure

This is an extension of the CIE1976 formula defined as (CIE, 1976)

$$\Delta E_{76} = \sqrt{(L_{\mathrm{ref}} - L_{\mathrm{test}})^2 + (a_{\mathrm{ref}} - a_{\mathrm{test}})^2 + (b_{\mathrm{ref}} - b_{\mathrm{test}})^2}, \tag{6.1}$$

where $(L_{\mathrm{ref}}, a_{\mathrm{ref}}, b_{\mathrm{ref}})$ and $(L_{\mathrm{test}}, a_{\mathrm{test}}, b_{\mathrm{test}})$ correspond to a given reference and test colors in the CIELAB color space, respectively. L is the lightness scale, the a axis corresponds to red-green opponent hues and the b axis corresponds to the yellow-blue opponent hues (CIE, 1976). The extension considers differences in chroma and the masking effect for computing CDs in images. Particularly, (Chou and Liu, 2007) defined the visibility of a just noticeable CD (JNCD) as

$$V = 2.3\alpha(\upsilon)\beta(\mu_Y, |\nabla Y|)S_C(a, b),$$

where $S_C(a, b) = 1 + 0.045\sqrt{a^2 + b^2}$ is the weighting function to adjust the dimension along the chroma axis, $\alpha(\upsilon) = \frac{\upsilon}{150} + 1$ is a scale function that models the increased tolerance of differences in non-uniform color patches. The

uniformity of the region is defined as the average variance of the three color components over a square area surrounding the pixel of interest, i.e., $\upsilon = \left(\sigma_L^2 + \sigma_a^2 + \sigma_b^2\right)/3$. $\beta(\mu_Y, |\nabla Y|) = \rho(\mu_Y)|\nabla Y| + 1$ is a scaling function that models the texture masking effect under different luminance levels, where

$$\rho(\mu_Y) = \begin{cases} 0.06 & \text{if } \mu_Y \leq 60 \\ 0.04 & \text{if } 60 < \mu_Y \leq 100 \\ 0.01 & \text{if } 100 < \mu_Y \leq 140 \\ 0.03 & \text{if } 140 < \mu_Y \leq 255 \end{cases},$$

μ_Y is the mean luminance value over a square area surrounding the pixel of interest and $|\nabla Y|$ is the magnitude of the spatial gradient. The overall CD measure is computed as the average value of the pixel differences using the following formula

$$\Delta E^{\mathrm{J}} = \Big((|\Delta L| - V)^2 \, \delta(|\Delta L|) + (|\Delta a| - V)^2 \, \delta(|\Delta a|) + (|\Delta b| - V)^2 \, \delta(|\Delta b|) \Big)^{1/2}, \quad (6.2)$$

where $\delta(x) = 1$ if $x > 0$ and $\delta(x) = 0$ otherwise.

The CIEDE2000 formula

The CIEDE2000 formula is a procedure introduced with the purpose of measuring just noticeable CDs between two given colors. Even though the CIEDE2000 formula was not specifically designed for computing CDs in natural scene color images, it is one of the most well-known CD formulas to date and it has shown better performance than other reported formulas for computing CDs in homogeneous color samples (Moroney et al., 2002; Habekost, 2013). The formula was designed using the outcome of psychovisual studies with both trained and untrained observers who were asked to judge CDs in homogeneous color samples. The CIEDE2000 formula is defined as follows:

$$\Delta E_{00} = \sqrt{\left(\frac{\Delta L}{k_L S_L}\right)^2 + \left(\frac{\Delta C}{k_C S_C}\right)^2 + \left(\frac{\Delta H}{k_H S_H}\right)^2 + R_T \frac{\Delta C}{k_C S_C} \frac{\Delta H}{k_H S_H}}. \quad (6.3)$$

The k_L, k_C, and k_H are correction factors related to the observation environment in terms of lightness (L), chroma (C) and hue (H), respectively. These weighting factors are usually set to the value of 1. For the rest of the terms in the formula and further details, we refer the reader to (Sharma, 2002). In general this measure is used pixel-wise resulting in a CD map. Then, the overall CD is computed by using the average value of such a map (Johnson and Fairchild, 2003).

Spatial extensions of the CIEDE2000 formula

A spatial extension of the CIEDE2000 formula was first proposed by (Zhang and Wandell, 1997) and further explored in other related works for measuring

CDs in images (Johnson and Fairchild, 2003; Zhang et al., 2010a; He et al., 2011). In (Zhang and Wandell, 1997) extension, image pairs are first converted into an opponent-color space approximating white-black, red-green, and yellow-blue color perceptions (Zhang and Wandell, 1997; Johnson et al., 2010). Thereafter, the images are filtered with approximations of the contrast sensitivity function to simulate the blur property of the human eyes. After the filtering, the images are transformed to the CIELAB color space for computing pixel-wise the CIDE2000 formula. In the following paragraph we briefly describe the spatial extension proposed by (Zhang et al., 2010a) with the purpose of illustrating its computation.

First the image is transformed into an opponent color space using the following linear transformation:

$$\begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix} = \begin{bmatrix} 27.3158 & 61.3096 & -1.8644 \\ -12.9955 & 2.7011 & -11.6044 \\ -8.2168 & -32.7442 & 44.2130 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix},$$

where O_1, O_2 and O_3 represent opponent color components of luminance (L), red-green (R−G) and blue-yellow (B−Y), respectively. The purpose of such a color transformation is to avoid mixing color components during the filtering process (Zhang and Wandell, 1997; Zhang et al., 2010a; He et al., 2011). Thereafter, each color component is independently filtered using a Gaussian function in the spatial domain to approximate the blur property of the human eyes. Afterwards, O_1, O_2 and O_3 values are transformed into CIELAB color space. Then, pixel-wise differences using the CIEDE2000 formula are computed for obtaining a CD map. Finally, the average value of the CD map is computed as overall image CD, termed, $\Delta E^{\mathrm{S}}_{00}$.

CD based on the Mahalanobis distance

(Imai et al., 2001) have proposed an alternative CD measure based on the Mahalanobis distance and the CIELAB color space. This measure uses the covariance between each color component as a weighting factor. These weighting factors were introduced with the purpose of considering the correlation between the CIELAB color components (Imai et al., 2001). In that work, the Mahalanobis distance is computed between two color pairs as follows:

$$\Delta E^{\mathrm{M}} = \sqrt{\begin{bmatrix} \Delta L & \Delta C & \Delta h \end{bmatrix} \begin{bmatrix} \sigma_{LL} & \sigma_{LC} & \sigma_{Lh} \\ \sigma_{CL} & \sigma_{CC} & \sigma_{Ch} \\ \sigma_{hL} & \sigma_{hC} & \sigma_{hh} \end{bmatrix}^{-1} \begin{bmatrix} \Delta L \\ \Delta C \\ \Delta h \end{bmatrix}}, \tag{6.4}$$

where $\sigma_{pq} = \frac{1}{n-1} \sum_{i=1}^{n} (p_i - \mu_p)(q_i - \mu_q)$ is the covariance between p and q. μ_p and μ_q are the average values of p and q, respectively. Here, p and q belong to the set $\{L, C, h\}$. L, C and h are the lightness, the chroma and the hue values defined in the CIELAB color space. This formula is applied pixel-wise and thereafter the average value is computed as the overall CD. Note that the covariance matrix is computed by using the entire reference image.

Colorfulness

Colorfulness is a color attribute first proposed by (Hasler and Susstrunk, 2003) with the purpose of measuring the color intensity (chromatic level) of natural scene color images. It is usually estimated using statistics of color components. In the following paragraphs we introduce the most well-known colorfulness measure. Given a set of values in RGB color space, colorfulness is computed by first transforming the RGB color components into a very simple opponent color space with two color components, i.e., $\alpha = R - G$, $\beta = 0.5(R + G) - B$. (Gao et al., 2013) proposed to introduce a logarithmic operation into the measure for simulating the logarithm sensation of the human visual system resulting in the following formula

$$\mathrm{Cf}^{\mathrm{G}} = 0.02 \log\left(\frac{\sigma_\alpha^2}{\mu_\alpha^{0.2}}\right) \log\left(\frac{\sigma_\beta^2}{\mu_\beta^{0.2}}\right). \tag{6.5}$$

where μ_α, μ_β, σ_α^2 and σ_β^2 are the means and standard deviations of α and β color components. Colorfulness values can be used to compute an overall CD between two given images by computing the difference of the obtained colorfulness values (Hasler and Susstrunk, 2003). For instance, $\Delta\mathrm{Cf}^{\mathrm{G}} = \mathrm{Cf}^{\mathrm{G}}_{\mathrm{ref}} - \mathrm{Cf}^{\mathrm{G}}_{\mathrm{test}}$ is a colorfulness difference where $\mathrm{Cf}^{\mathrm{G}}_{\mathrm{ref}}$ and $\mathrm{Cf}^{\mathrm{G}}_{\mathrm{test}}$ are the colorfulness measure computed using Equation (6.5) on the reference and test images, respectively.

Color extension of the structural similarity index

A color extension of the structural similarity index (SSIM) has been proposed by (Toet and Lucassen, 2003). The authors based their measure on the fact that the human visual system processes images in three uncorrelated color components: one luminance and two opponent color components. Thus, each color component will contribute independently to perceived image differences, and should therefore be calculated independently before combining them into an overall difference (Toet and Lucassen, 2003; Hassan and Bhagvati, 2012). That is, given two images in RGB color space, first both images are transformed into a perceptually uncorrelated color space by using the following formulas (Toet and Lucassen, 2003)

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.3811 & 0.5783 & 0.0402 \\ 0.1967 & 0.7244 & 0.0782 \\ 0.0241 & 0.1288 & 0.8444 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix},$$

$$\begin{bmatrix} \ell \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1/\sqrt{6} & 0 \\ 0 & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \log(L) \\ \log(M) \\ \log(S) \end{bmatrix}.$$

Second, the SSIM is independently computed on ℓ, α and β color components as discussed in (Wang and Bovik, 2002). Third, the CD measure (color SSIM) is computed as

$$\mathrm{CSSIM} = \sqrt{w_\ell \mathrm{SSIM}_\ell^2 + w_\alpha \mathrm{SSIM}_\alpha^2 + w_\beta \mathrm{SSIM}_\beta^2}, \tag{6.6}$$

where SSIM_ℓ, SSIM_α and SSIM_β are the structural similarity indices computed on ℓ, α and β color components between the reference and test images. The weighting factors were found experimentally ($w_\ell = 3.05$, $w_\alpha = 1.1$ and $w_\beta = 0.85$) (Toet and Lucassen, 2003).

Chroma spread and extreme

Chroma spread and chroma extreme are two CD indices proposed by (Pinson and Wolf, 2004a) with the purpose of quantifying, respectively, changes in the spread of the distribution of two-dimensional color samples and severe localized color impairments. Given two image samples in $\mathrm{YC_BC_R}$ color space (cf. ITU-R BT.601-5 recommendation (ITU, 1995)), the chroma spread and extreme measures are computed as follows:

- Chroma spread ($\mathrm{Ch_s}$):
  1. Divide the $\mathrm{C_B}$ and $\mathrm{C_R}$ color components into separate regions of 8×8 pixels.
  2. Compute the mean of each region.
  3. Compare the reference and processed means obtained in step 2 using Euclidean distance.
  4. Spatially collapse by computing the standard deviation of the resulting differences.
- Chroma extreme ($\mathrm{Ch_e}$):
  1. Perform steps 1 through 3 from chroma spread.
  2. Spatially collapse by computing the average of the worst 1% and subtract from it the 99% level.

Thereafter, the two measures are weighted summed to get an overall CD

$$\mathrm{Ch} = \omega_s \mathrm{Ch_s} + \omega_e \mathrm{Ch_e}, \tag{6.7}$$

where $\omega_s = 0.0192$ and $\omega_e = 0.0076$ were obtained empirically using training samples from the VQEG FR-TV Phase II database (Pinson and Wolf, 2004a).

CDs based on histogram intersection

A technique known as histogram intersection has been widely studied and it is considered to be effective for color-image indexing. The key issue of this algorithm is the selection of an appropriate color space and an optimal quantization of the selected color space. Particularly, (Lee et al., 2005) have studied the performance of various color spaces and quantization steps in various images for identifying those with higher agreement with human judgment of image similarity measurement. The authors found, after exploring six color spaces and twelve quantization levels, that the CIELAB color space generally performs

better than the other color spaces in most conditions for most of the considered quantization levels (Lee et al., 2005). Additionally, the authors concluded that a 512 bins histogram (8 bins per each of the CIELAB color components) produces accurate results and further increase in the number of bins brings negligible improvement. The CD measure is computed using the following two steps. First, a 512 bins color histogram is computed on the reference and the test images. Second, the intersection is computed between the color histograms f_{ref} and f_{test} as

$$K_{\bigcap} = \sum_{i=1}^{8}\sum_{j=1}^{8}\sum_{k=1}^{8} \min\left(f_{\text{ref}}(i,j,k), f_{\text{test}}(i,j,k)\right). \tag{6.8}$$

Weighted CIEDE2000

This is an extension of the CIEDE2000 formula for evaluating CDs in images. It has been derived based on observations over cases where CIEDE2000 formula fails. The analysis revealed that the lightness, chroma and hue weighting factors need to be modified for natural scene color images and could be image dependent. The weighted CIEDE2000 is based on the fact that CDs in larger areas of the same color should be weighted higher because human eyes tend to be more tolerant towards CD in smaller areas (Hong and Luo, 2006). Then, assuming the images in CIELAB color space, (Hong and Luo, 2006) have proposed the following procedure:

1. Compute the CIEDE2000 formula pixel by pixel (cf. Equation (6.3)).
2. Extract from the hue angle of the reference sample a 180 bins histogram (f_h).
3. Sort the normalized histogram ($\sum f_h = 1$) in ascending order.
4. Assign the following weights to the sorted histogram

$$f_{h_{\text{sorted}}}(k) = \begin{cases} 0.25 * f_{h_{\text{sorted}}}(k) & \text{if } 1 \le k < n_{25\%} \\ 0.5 * f_{h_{\text{sorted}}}(k) & \text{if } n_{25\%} \le k < n_{50\%} \\ f_{h_{\text{sorted}}}(k) & \text{if } n_{50\%} \le k < n_{75\%} \\ 2.25 * f_{h_{\text{sorted}}}(k) & \text{otherwise} \end{cases},$$

   where $n_{25\%}$, $n_{50\%}$ and $n_{75\%}$ are selected such that the sorted histogram is divided into quartiles, i.e., it is divided in four equal groups ($\sum_{k=1}^{n_{25\%}} f_{h_{\text{sorted}}}(k) = 0.25$, $\sum_{k=n_{25\%}}^{n_{50\%}} f_{h_{\text{sorted}}}(k) = 0.25$, $\sum_{k=n_{50\%}}^{n_{75\%}} f_{h_{\text{sorted}}}(k) = 0.25$ and $\sum_{k=n_{75\%}}^{n_{100\%}} f_{h_{\text{sorted}}}(k) = 0.25$).
5. For each hue angle bin, compute the mean value of the CIEDE2000 CDs having that same hue angle bin. That is,

$$\mathbf{e}(\mathrm{h}) = \frac{\sum_{i,j\in\Omega^{\mathrm{h}}} \Delta E_{00}(i,j)}{|\Omega^{\mathrm{h}}|},$$

where $\Delta E_{00}(i,j)$ is the (i,j)th CD value computed between the two given image samples. Here, $\Delta E_{00}(i,j) \in \Omega^{\mathrm{h}}$ if the assigned bin to h in the (i,j)th pixel is equal to h., i.e., $h(i,j) = \mathrm{h}$. $|\Omega|$ is the number of elements in the set Ω.

6. Sort $\mathbf{e}$ by using the collection of indices that describes the arrangement of the elements of f_h into $f_{h_{\mathrm{sorted}}}$ (cf. step 3).

7. The overall CD is computed with the following formula

$$\Delta E_{00}^{\omega} = \sum_{k=1}^{180} f_{h_{\mathrm{sorted}}}(k)\mathbf{e}^2_{\mathrm{sorted}}(k). \tag{6.9}$$

Image CD measure based on image appearance models

Image appearance models have been developed over the recent years as a tool to predict perceived changes between different types of imaging systems, e.g., they can be used to generate a visual match between a hardcopy print and a softcopy display. Unlike the color appearance models, the image appearance models also includes attributes of perception of contrast, graininess and sharpness (Johnson, 2006). For instance, the most well-known image appearance model is the CIECAM02 developed by the CIE (CIE, 2004). This model was generated as an alternative to the CIELAB color appearance model to include mechanisms to account for changes in overall luminance, background luminance and surrounding viewing conditions (Johnson, 2006). Further details on the evolution of image appearance models can be found in (Fairchild, 2013). In this thesis, we describe the standard image appearance model recommended in the CIE-159:2004 technical report (CIE, 2004) which is used in the so called image color appearance model (iCAM) (Johnson, 2006; Fairchild, 2013).

First, the two images under consideration are filtered by using a 2D contrast sensitivity function in an opponent color space termed $\mathrm{YC_1C_2}$ (Johnson, 2006). Thereafter, a chromatic adaptation is applied on both images by using the method proposed in the CIECAM02 (CIE, 2004). Then, a local contrast predictor based on a low-pass version of the luminance component is used to estimate simultaneous contrast changes between the images (Fairchild and Johnson, 2004). Afterwards, a transformation into a uniform color space (cf. IPT color appearance model (Johnson et al., 2010)) is used for computing pixel-wise CDs resulting in a CD map. Finally, the overall CD measure, termed ΔE^{I}, is computed as the average of the CD map.

CD based on OSA-UCS color appearance model

In this approach CDs are measured by using an Euclidean based formula for small-medium CDs in the log-compressed OSA-UCS color space (Huertas et al., 2006; Oleari et al., 2008). The space was built with the purpose of alleviating the inadequacy of uniform color spaces to account for large CDs (Huertas et al., 2006). Given the reference and test images in the CIE xyY and XYZ

color spaces (cf. Smith and Guild (Smith and Guild, 1931)), the OSA-UCS coordinates representing luminance, chroma and hue, respectively, are computed as follows:

$$L^{\mathrm{OSA}} = \frac{1}{\sqrt{2}}\left(5.9\left(\left(Y_0^{1/3} - \frac{2}{3}\right) + 0.042\left(Y_0 - 30\right)^{1/3}\right) - 14.4\right),$$

$$C^{\mathrm{OSA}} = \sqrt{J^2 + G^2},$$

$$H^{\mathrm{OSA}} = \arctan\left(\frac{J}{-G}\right),$$

with $Y_0 = \mathrm{Y}\left(4.4934\mathrm{x}^2 + 4.3034\mathrm{y}^2 - 4.276\mathrm{xy} - 1.3744\mathrm{x} - 2.5643\mathrm{y} + 1.8103\right)$,

$$\begin{bmatrix} J \\ G \end{bmatrix} = \begin{bmatrix} 1.147L^{\mathrm{OSA}} + 14.178 & 0 \\ 0 & -(1.528L^{\mathrm{OSA}} + 18,504) \end{bmatrix} \begin{bmatrix} 0.1792 & 0.9837 \\ 0.9482 & -0.3175 \end{bmatrix} \begin{bmatrix} \ln\left(\frac{A/B}{0.9366}\right) \\ \ln\left(\frac{B/C}{0.9807}\right) \end{bmatrix} \tag{6.10}$$

and

$$\begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} 0.6597 & 0.4492 & -0.1089 \\ -0.3053 & 1.2126 & 0.0927 \\ -0.0374 & 0.4795 & 0.5579 \end{bmatrix} \begin{bmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \end{bmatrix}.$$

The OSA-UCS difference measure is defined as (Huertas et al., 2006)

$$\Delta E^{\mathrm{O}} = 10\sqrt{\left(\frac{\Delta L^{\mathrm{OSA}}}{k_L S_L}\right)^2 + \left(\frac{\Delta C^{\mathrm{OSA}}}{k_C S_C}\right)^2 + \left(\frac{\Delta H^{\mathrm{OSA}}}{k_H S_H}\right)^2}, \tag{6.11}$$

where k_L, k_C as well as k_S are typically set to 1.

$$S_L = 2.499 + 0.07\bar{L}^{\mathrm{OSA}},$$

$$S_C = 1.235 + 0.58\bar{C}^{\mathrm{OSA}} \quad \text{and}$$

$$S_H = 1.392 + 0.17\bar{H}^{\mathrm{OSA}}$$

where found experimentally such that ΔE^{O} agrees with perceived CDs between homogeneous color samples. Here,

$$\bar{L}^{\mathrm{OSA}} = \frac{L^{\mathrm{OSA}}_{\mathrm{ref}} + L^{\mathrm{OSA}}_{\mathrm{test}}}{2},$$

$$\bar{C}^{\mathrm{OSA}} = \frac{C^{\mathrm{OSA}}_{\mathrm{ref}} + C^{\mathrm{OSA}}_{\mathrm{test}}}{2} \quad \text{and}$$

$$\bar{H}^{\mathrm{OSA}} = \frac{H^{\mathrm{OSA}}_{\mathrm{ref}} + H^{\mathrm{OSA}}_{\mathrm{test}}}{2}$$

are the average coordinates between reference and test samples. This measure is applied pixel-wise resulting in a CD map which is later averaged to obtain the overall CD between two images.

Spatial extension of the OSA-UCS based CD

This CD measure proposed by (Simone et al., 2009) uses the same spatial processing performed in the spatial extension of CIEDE2000 formula ($\Delta E^{\mathrm{S}}_{00}$). Thereafter, the OSA-UCS based CD measure is applied pixel by pixel to the resulting images (cf. Equation (6.11)). The overall CD measure, termed ΔE^{SO}, is obtained by averaging all CDs.

CD measure based on local spatial differences

(Ouni et al., 2008) have proposed an extension of the CIEDE2000 formula to take into account neighbour pixels in the CD computation. First the CIEDE2000 formula is applied pixel by pixel between the two given images (cf. Equation (6.3)). Afterwards, the obtained pixel-wise differences are filtered by using the following kernel that takes into account the distance between the central pixel and its neighbours

$$W = \begin{bmatrix} 0.5 & 1 & 0.5 \\ 1 & 0 & 1 \\ 0.5 & 1 & 0.5 \end{bmatrix}.$$

This results in an image that takes into account the spatial processing of the visual system, termed $\Delta \bar{E}_{00}$ (Ouni et al., 2008). Thereafter, the weighted pixel differences are added to the CIEDE2000 values and later normalized, i.e.,

$$\Delta E^{D} = \frac{1}{7}\left(\Delta E_{00} + \Delta \bar{E}_{00}\right). \tag{6.12}$$

Finally, the average over the pixel differences is computed as overall CD.

Image CD measure on Hue and Saturation

A reduce reference CD measure based on the HSI color space, cf. (Smith, 1978), has been proposed by (Ming et al., 2009). The CD measure is obtained by the linear combination of two values:

$$\Delta E^{HS} = w_H \Delta \mu_H + w_S \Delta \mu_S, \tag{6.13}$$

where $\Delta \mu_H = |\mu_{H_{\mathrm{ref}}} - \mu_{H_{\mathrm{test}}}|$ and $\Delta \mu_S = |\mu_{S_{\mathrm{ref}}} - \mu_{S_{\mathrm{test}}}|$. Here, $\mu_{H_{\mathrm{ref}}}$, $\mu_{H_{\mathrm{test}}}$, $\mu_{S_{\mathrm{ref}}}$ and $\mu_{S_{\mathrm{test}}}$ are spatial averages of hue and saturation components for a given pair of color images in HSI color space. The weights ($\omega_H = 0.3$ and $\omega_S = 0.1$) were found experimentally such that ΔE_{HS} correlates well with the actual perceived CD (Ming et al., 2009).

Adaptive spatio-chromatic image difference

In this framework CDs are computed based on an adaptive signal decomposition method (Rajashekar et al., 2009, 2010). This method decomposes local blocks of the differences between the reference and the test image using a set

of basis functions adapted to the data of the reference. The adaptive functions are chosen to capture differences in luminance, hue and saturation. The CD measure is defined as

$$\Delta E^{\mathrm{A}} = (W_M^2 + M^T M)^{-1}(M\Delta\mathbf{x}), \tag{6.14}$$

where

$$W_M^2 = \mathbb{I}_{6\times 6}\begin{bmatrix}0.01\\0.01\\0.01\\0.01\\0.01\\0.36\end{bmatrix}, \Delta\mathbf{x} = \begin{bmatrix} r_{\mathrm{ref},1} - r_{\mathrm{test},1}\\ g_{\mathrm{ref},1} - g_{\mathrm{test},1}\\ b_{\mathrm{ref},1} - b_{\mathrm{test},1}\\ \vdots \\ r_{\mathrm{ref},n} - r_{\mathrm{test},n}\\ g_{\mathrm{ref},n} - g_{\mathrm{test},n}\\ b_{\mathrm{ref},n} - b_{\mathrm{test},n}\end{bmatrix}.$$

$(r_{\mathrm{ref},j}, g_{\mathrm{ref},j}, b_{\mathrm{ref},j})$ and $(r_{\mathrm{test},j}, g_{\mathrm{test},j}, b_{\mathrm{test},j})$ are the given pair of color values for the jth pixel in a local block of n pixels. $\mathbb{I}_{6\times 6}$ is the identity matrix of size 6×6. The adaptive basis is defined as $M = [\mathbf{m}_1, \ldots, \mathbf{m}_6]$, where each column vector account for one of the following attributes: changes on white balance $(\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3)$, luminance $(\mathbf{m}_4)$, chroma $(\mathbf{m}_5)$ and hue $(\mathbf{m}_6)$. Each vector is defined as follows:

$$\mathbf{m}_1 = \begin{bmatrix} r_{\mathrm{ref},1}\\0\\0\\ \vdots \\ r_{\mathrm{ref},n}\\0\\0\end{bmatrix}, \mathbf{m}_2 = \begin{bmatrix} 0\\ g_{\mathrm{ref},1}\\0\\ \vdots \\0\\ g_{\mathrm{ref},n}\\0\end{bmatrix}, \mathbf{m}_3 = \begin{bmatrix} 0\\0\\ b_{\mathrm{ref},1}\\ \vdots \\0\\0\\ b_{\mathrm{ref},n}\end{bmatrix}$$

$$\mathbf{m}_4 = \begin{bmatrix} l_{\mathrm{ref},1}\\ l_{\mathrm{ref},1}\\ l_{\mathrm{ref},1}\\ \vdots \\ l_{\mathrm{ref},n}\\ l_{\mathrm{ref},n}\\ l_{\mathrm{ref},n}\end{bmatrix}, \mathbf{m}_5 = \begin{bmatrix} r_{\mathrm{ref},1}\\ g_{\mathrm{ref},1}\\ b_{\mathrm{ref},1}\\ \vdots \\ r_{\mathrm{ref},n}\\ g_{\mathrm{ref},n}\\ b_{\mathrm{ref},n}\end{bmatrix} - \mathbf{m}_4, \text{and, } \mathbf{m}_6 = \begin{bmatrix}\mathbf{h}_1\\ \vdots \\ \mathbf{h}_n\end{bmatrix},$$

where $l_{\mathrm{ref},j} = (r_{\mathrm{ref},j} + g_{\mathrm{ref},j} + b_{\mathrm{ref},j})/3$ and $\mathbf{h}_j = l_{\mathrm{ref},j}[b_{\mathrm{ref},j} - g_{\mathrm{ref},j}, r_{\mathrm{ref},j} - b_{\mathrm{ref},j}, g_{\mathrm{ref},j} - r_{\mathrm{ref},j}]^T$. This procedure is repeated pixel by pixel by using a 3×3 sliding window resulting in a CD map. Thereafter, the overall CD measure is computed as the average of the CD map.

Spatial hue angle metric (SHAME)

The SHAME measure has been proposed by (Pedersen and Hardeberg, 2009). It combines the weighted CIEDE2000 formula proposed by (Hong and Luo,

2006) with additional spatial processing which takes into account the spatial properties of the human visual system. The weighted CIEDE2000 measure has been selected as the basis because it corrects some of the drawbacks of the CIEDE2000 formula, such as the uneven weights of the different hue angles in the overall CD computation. The CD measure is computed as follows: first a spatial processing on each color component is applied to the reference and the test images by using the same filters and opponent color space as the spatial extension of the CIEDE2000 formula ($\Delta E^{\mathrm{S}}_{00}$). Afterwards, the resulting filtered images are used as input to the weighted CIEDE2000 measure (ΔE^{ω}_{00}) to obtain the overall CD, termed ΔE^{SH}.

Color image difference

The color image difference is a measure based on the hypothesis that the human visual system is sensitive to lightness, chroma, and hue differences. In the measure, three indices are computed by using local statistics of differences of lightness, chroma and hue (Lissner et al., 2013; Preiss et al., 2014). The color image difference is computed from two images in the CIELAB color space as follows: let

$$\Delta L = L_{\mathrm{ref}} - L_{\mathrm{test}}, \quad \Delta C = C_{\mathrm{ref}} - C_{\mathrm{test}} \quad \text{and}$$
$$\Delta H = \sqrt{(a_{\mathrm{ref}} - a_{\mathrm{test}})^2 + (b_{\mathrm{ref}} - b_{\mathrm{test}})^2 - \Delta C^2}$$

be the lightness, chroma and hue differences computed pixel-wise after filtering independently the three color components using a Gaussian filter, cf. (Lissner et al., 2013). Then, the three overall indices, termed l_L, l_C and l_H, are computed as the average of ΔL^-, ΔC^- and ΔH^-, respectively. Where

$$\Delta L^- = 1 - \frac{1}{0.002\Delta L^2 + 1}, \quad \Delta C^- = 1 - \frac{1}{0.002\Delta C^2 + 1} \quad \text{and}$$
$$\Delta H^- = 1 - \frac{1}{0.008\Delta H^2 + 1}.$$

Finally, the overall color image difference is compute as the product between l_L, l_C and l_H, i.e.,

$$\Delta E^{\mathrm{CI}} = 1 - l_L l_C l_H. \tag{6.15}$$

Image CD measure based on circular hue

This measure quantifies the differences between local lightness, hue and chroma information between two color samples. (Lee and Rogers, 2014) have proposed to compare hue information based on the theory of circular statistics to take into account the periodicity of the hue component. In particular, images are spatially processed by using the same schema of the spatial extension of the CIEDE2000 formula ($\Delta E^{\mathrm{S}}_{00}$). Afterwards, the hue component of the given image pair are compared as

$$\Delta \bar{H} = \frac{2\bar{H}_{\mathrm{ref}}\bar{H}_{\mathrm{test}} + 12.96}{\bar{H}^2_{\mathrm{ref}} + \bar{H}^2_{\mathrm{test}} + 12.96}.$$

The circular mean ($\bar{H}$) is computed from a set of n neighbour pixels in the hue component as:

$$\bar{H} = \arctan\left(\frac{\sum_{j=1}^{n}\cos(H_j)}{n}, \frac{\sum_{j=1}^{n}\sin(H_j)}{n}\right).$$

Chroma comparison is performed using arithmetic means ($\bar{C}$) in the same set of pixel values, i.e.,

$$\Delta\bar{C} = \frac{2\bar{C}_{\text{ref}}\bar{C}_{\text{test}} + 3.24}{\bar{C}^2_{\text{ref}} + \bar{C}^2_{\text{test}} + 3.24}.$$

Lightness comparison ($\Delta\bar{L}$) is computed as the SSIM between the given images on the lightness component. All indices are computed using a sliding window to get a CD map per color component. Thereafter, the CD map is computed as the product of the independent color component differences, i.e.,

$$\Delta E^{CH} = 1 - \Delta\bar{H}\Delta\bar{C}\Delta\bar{L}. \tag{6.16}$$

Finally, the overall CD is computed as the average of the CD map.

6.2.2 Summary

We have explored eighteen color difference measures listed in Table 6.1. The symbol is the notation used in this work for referring to a specific CD measure. Color space is the color space or appearance model used for computing the CDs. SP (Spatial processing) is whether or not neighboring pixels are taken into account in computing the CD measure. Overall CD describes the technique for computing the overall CD measure using the pixel-wise differences.

In general, we have found: eight extensions of the CIEDE2000, four based on statistics of color components, two extensions of the SSIM and four based on other color appearance models. The explored measures use 8 CAMs: CIELAB (used by 11 out of 18 measures), 2-component opponent color space (OCC) (1), OSA-UCS (2), $\ell\alpha\beta$ (1), $\text{YC}_\text{B}\text{C}_\text{R}$ (1), HSI (1), IPT (1) and RGB (1). For more information about these CAMs, the reader is referred to the original publications listed in Table 6.1. Note that the CIELAB appearance model is the most popular CAM for computing CDs in natural scene color images. 9 out of 18 measures do not consider any spatial processing. Finally, irrespective of whether the measure has spatial processing or not, the overall difference in 11 out of the 18 CD measures is computed as the average of the pixel-wise differences.

Traditionally, computing CDs in images has been accomplished by using a CD formula on a pixel-by-pixel basis and then examining statistics such as mean, median or maximum. However, subjective evaluation of perceived color differences has shown that, when observing a color image, the observer makes the color sensation from a number of pixels and not a single pixel color (Liu et al., 2010). Also, the studies in color enhancement have shown that the perceived color by a human depends on the amount of spatial variation and texture

Table 6.1: State-of-the-art summary studied in this Chapter.

Measure name	Symbol	Color space	SP	Overall CD
CIEDE2000 formula (Luo et al., 2001)	ΔE_{00}	CIELAB (CIE, 1976)	No	Average of pixel-wise CDs
Spatial extension CIEDE2000 (Zhang and Wandell, 1997)	$\Delta E_{00}^{\mathrm{S}}$	CIELAB (CIE, 1976)	Yes	Average of pixel-wise CDs
CD based on Mahalanobis distance (Imai et al., 2001)	ΔE^{M}	CIELAB (CIE, 1976)	No	Average of pixel-wise CDs
Colorfulness (Gao et al., 2013)	$\Delta \mathrm{Cf}^{\mathrm{G}}$	2-component OCC (Hasler and Susstrunk, 2003)	No	Difference in global descriptive statistics of color components
Color extension of the SSIM (Toet and Lucassen, 2003)	CSSIM	$\ell\alpha\beta$ (Ruderman et al., 1998)	Yes	Average of pixel-wise CDs
Chroma spread and extreme (Pinson and Wolf, 2004a)	Ch	$\mathrm{YC_BC_R}$ (ITU, 1995)	Yes	Statistics of local differences between color features
CD based on histogram intersection (Lee et al., 2005)	$K_{\bigcap}$	CIELAB (CIE, 1976)	No	Histogram intersection of color histograms
Weighted CIEDE2000 (Hong and Luo, 2006)	ΔE_{00}^{ω}	CIELAB (CIE, 1976)	No	Weighted average of pixel-wise CDs
Image CD based on CAM (Johnson, 2006)	ΔE^{I}	IPT (Johnson, 2006)	Yes	Average of pixel-wise CDs
CD based on OSA-UCS (Huertas et al., 2006)	ΔE^{O}	Log-compressed OSA-UCS (Oleari, 2004)	No	Average of pixel-wise CDs
Spatial extension OSA-UCS CD (Simone et al., 2009)	ΔE^{SO}	Log-compressed OSA-UCS (Oleari, 2004)	Yes	Average of pixel-wise CDs
Just noticeable CD measure (Chou and Liu, 2007)	ΔE^{J}	CIELAB (CIE, 1976)	Yes	Weighted Average of pixel-wise CDs
CD based on local spatial differences (Ouni et al., 2008)	ΔE^{D}	CIELAB (CIE, 1976)	Yes	Average of pixel-wise CDs
Image CD on Hue and Saturation (Ming et al., 2009)	ΔE^{HS}	HSI (Smith, 1978)	No	Difference in global descriptive statistics of color components
Adaptive spatio-chromatic image difference (Rajashekar et al., 2009)	ΔE^{A}	RGB (Sharma, 2002)	Yes	Average of pixel-wise CDs
Spatial hue angle metric (Pedersen and Hardeberg, 2009)	ΔE^{SH}	CIELAB (CIE, 1976)	Yes	Weighted average of pixel-wise CDs
Color image difference (Lissner et al., 2013)	ΔE^{CI}	CIELAB (CIE, 1976)	Yes	Average of pixel-wise CDs
Image CD based on circular hue (Lee and Rogers, 2014)	ΔE^{CH}	CIELAB (CIE, 1976)	Yes	Average of pixel-wise CDs

in the scene (Palma-Amestoy et al., 2009; Bertalmio et al., 2009). That is, two image patches can be perceived by a human as the same color only under the same spatial distribution of pixel color values. Additionally, the experiments carried out in (Bando et al., 2005; Liu et al., 2010, 2013a) comparing color image differences showed that the observers tend to focus on certain areas of an image, usually, homogeneous areas or areas with the same texture pattern, and give their judgments mainly based on the color difference of those areas.

These findings show that the pixel-wise CDs between two images do not represent the CD sensation perceived by a human observer and human observers judge CD in natural scene color images based on the comparison of image patches with similar texture pattern. Note that the state-of-the-art CD

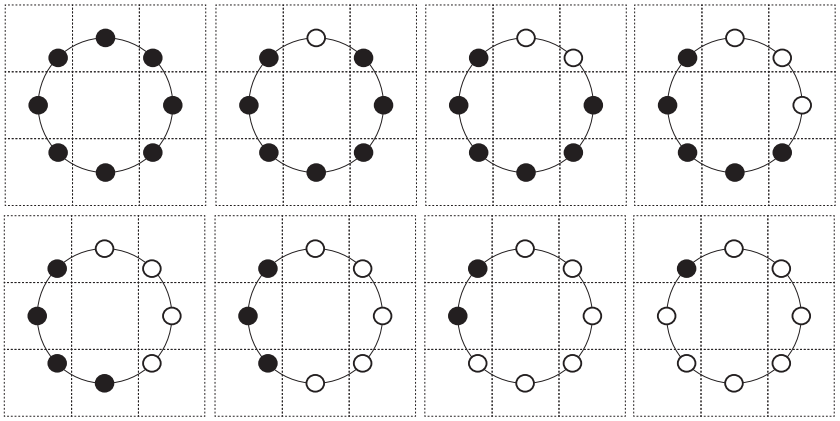

Figure 6.1: Texture primitives detected by the uLBP. Black points correspond to the binary value 0 while white points to 1.

measures do not consider the texture of the image in the CD computation.

6.3 Proposed method

In search for an adequate solution of the problem of computing color differences in natural scene color images, we propose a measure based on the fact that humans assess the differences in image color by comparing small image patches of similar texture. Therefore, we first look for an appropriate method to divide the image in patches with unique texture patterns to later compute the CDs on the obtained patches.

One common way of dividing an image into unique texture patterns is by using the well-known texture descriptors: the Local Binary Patterns (LBP). This method computes relative intensity relations between the pixels in a small neighborhood. See (Maenpaa, 2003) and Chapter 5 for details about this texture analysis technique. In particular, experimental results over all possible LBP patterns have shown that the subset called "uniform" LBP (uLBP), introduced in (Topi et al., 2000), covers 90% of all patterns in natural scene images (Topi et al., 2000; Fehr and Burkhardt, 2008). A LBP pattern is called uniform if the pattern contains at most two $0-1$ or $1-0$ transitions. Figure 6.1 shows the texture primitives detected by the uLBP. The black points correspond to the binary value 0 and the white points to 1. Note that any other texture primitive can be obtained by rotating or complementing the binary primitives shown in Figure 6.1.

Figure 6.2 shows examples of texture primitives computed using the uLBP. In the top we show the sample images while in the middle their corresponding uLBP primitives. In the bottom we show all the textured patches equal to the first texture primitive from Figure 6.1. The encircled patches in Figure 6.2

Figure 6.2: Example of texture primitives detected using uLBP. (top) sample image, (middle) uLBP primitives, (bottom) homogeneous patches for the first (top left corner) texture primitive from Figure 6.1. The encircled patches are examples of what we call homogeneous textured patches, i.e., a connected set of pixels with unique texture pattern.

are examples of what we call homogeneous textured patch, a set of connected pixels with an unique uLBP texture pattern.

After dividing the image into a set of unique texture patches using the uLBP descriptors, we are ready to perform the color comparison independently in each homogeneous textured patch. In this case, we can use one of the image CD indices explored in Section 6.2. Particularly, the statistics used in chroma spread and chroma extreme CD indices proposed by Pinson and Wolf (Pinson and Wolf, 2004a) have shown to be good measures of the change of spread in the color distribution and severe color differences, respectively. Accordingly, we propose to measure the CDs in the resulting homogeneous textured patches using the linear combination of the chroma spread and chroma extreme indices because they capture color distribution parameters relevant to the humans (Pinson and Wolf, 2004a). For computing the differences in the intensity channel, we use the well-known structural similarity index measure (SSIM) (Zhou et al., 2014).

Figure 6.3 shows the block diagram of the proposed methodology for computing color differences in natural scene color images. The computation of the proposed CD measure is summarized as follows.

1. The Reference and Test images are compared using the Euclidean distance of their corresponding C_B and C_R color components as well as using the SSIM between intensity components.

2. The uLBP is computed from the reference image to obtain the set of homogeneous textured patches (uLBP segmentation in Figure 6.3).

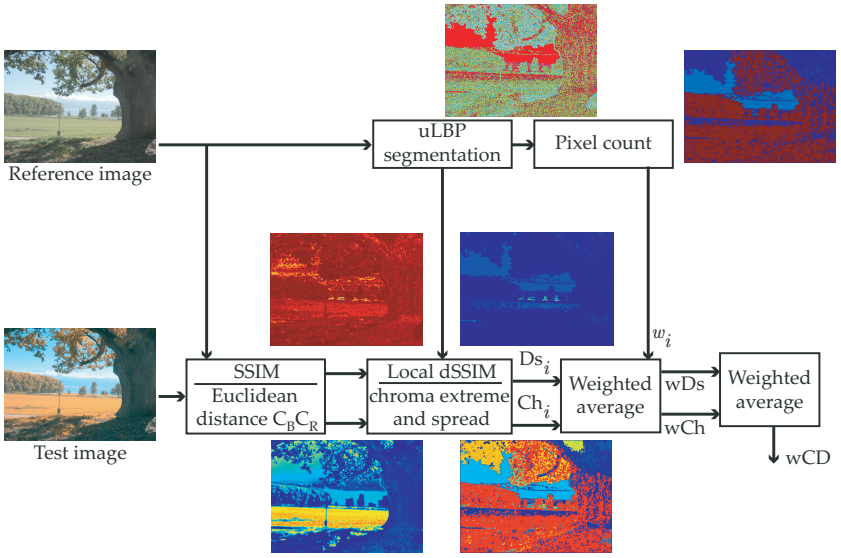

Figure 6.3: Block diagram of the proposed image CD measure.

3. In the *Local dSSIM, chroma extreme and spread* block, we compute for each homogeneous textured patch the chroma spread as the standard deviation of the resulting differences and the chroma extreme as the average of the worst 1% and subtract from it the 99% level (Pinson and Wolf, 2004a). Both indices are combined as the chroma spread-extreme index $\mathrm{Ch_i} = 0.0192\mathrm{Ch_s} + 0.0076\mathrm{Ch_e}$, for the ith homogeneous textured patch (Pinson and Wolf, 2004a). The linear combination was obtained empirically by (Pinson and Wolf, 2004a) using training samples from the VQEG FR-TV Phase II database. Similarly, we compute for each homogeneous textured patch the average value of the SSIM after being transformed to dissimilarity, i.e., $\mathrm{Ds}_i = \frac{1-\overline{\mathrm{SSIM}}_i}{2}$, where $\overline{\mathrm{SSIM}}_i$ is the average SSIM of the ith homogeneous textured patch.

4. The number of pixels in each homogeneous textured patch is count to be used as weights for the spatial pooling. The weights are computed as follows $w_i = \frac{n_i}{NM}$ where n_i is the number of pixels in the ith homogeneous textured patch, N and M are the number of rows and columns of the image, respectively. This assumption agrees with the well-known fact that human eyes tend to be more tolerant towards color difference of smaller image areas (Bando et al., 2005).

5. The global image color difference is computed as the weighted average of

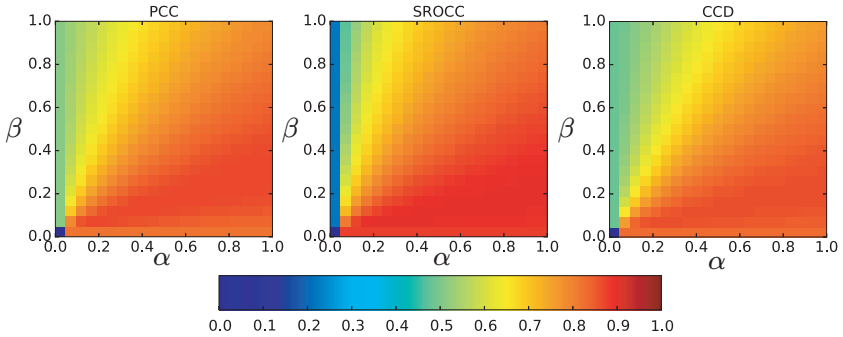

Figure 6.4: Performance of the proposed CD measure appraised on the test data of TID2013 database in function of the parameters α and β. Performance is given in terms of the PCC, the SROCC and CCD between the resulting CD measure and the corresponding subjective scores. The color bar represents the strength of the correlation from 0 to 1.

the resulting color differences per patch as

$$\text{wCh} = \sum_{i=1}^{K} w_i \text{Ch}_i,$$

$$\text{wDs} = \sum_{i=1}^{K} w_i \text{Ds}_i,$$

where Ch_i, Ds_i and w_i are the chroma spread-extreme index, the average dissimilarity index and the weight of the ith homogeneous textured patch for K patches, respectively. Note that the number of homogeneous textured patches (K) depends on the image content at hand. For instance, we have found (from left to right) 4458, 2788, 3658, 3828 and 3652 homogeneous textured patches in the images from Figure 6.2.

Finally, the global CD is computed as the weighted average of the two differences as follows

$$\text{wCD} = \alpha \text{wCh} + \beta \text{wDs}, \tag{6.17}$$

where α and β are weights that can be adjusted according to the application. In this case, since we are interested in evaluating color differences we give more importance to the color component, i.e., empirically we select the following weights: $\alpha = 0.7$ and $\beta = 0.3$.

Figure 6.4 shows the correlation between the humans scores in the test data of TID2013 database (see Section 6.4.1) and the proposed methodology in function of the parameters α and β. The highest correlation is achieved around the region of the selected parameter values ($\alpha = 0.7$ and $\beta = 0.3$). Also note that the performance decreases when a higher weight is assigned to the

differences computed in the intensity component of the image. Additionally, this experiment shows that it is possible to further investigate and tune α and β for different applications according to the importance of the differences in individual color components.

6.4 Results and Discussion

In this Section we describe the used test images and the performance comparison with the state-of-the-art measures. The performance comparison is made in terms of correlation indices computed between the CD measures and the subjective scores, which are considered as ground truth. The value of 1 indicates high correlation and 0 is no correlation between the tested CD measure and the subjective scores. Since the PCC, the SROCC and the CCD values obtained in this Chapter lead to analogous conclusions, we only describe our results in terms of the CCD but the analysis applies for all (PCC and SROCC) unless we indicate the opposite. We use the rule of the thumb for interpreting the size of a correlation coefficient (Mukaka, 2012) (see Section 2.3)

6.4.1 Test data

In order to carry out a meaningful performance analysis, in this Chapter the test data was selected to include the types of color alterations relevant for the most common applications considering CDs: color correction (Fezza et al., 2014; Ly et al., 2015), color quantization (Brun and Tremeau, 2002), color mapping (Morovic, 2008), color image similarity and retrieval (Mojsilovic et al., 2002). The output images in such tasks are typically affected by color distortions such as quantization noise, intensity shift, contrast change, change in color saturation and change in color balance (Sharma, 2002; Baranczuk et al., 2010; Wei and Mulligan, 2010; Fezza et al., 2014). The considered dataset was obtained from one publicly available image quality database named TID2013 described in the following paragraphs (see (Ponomarenko et al., 2015) and Appendix A.1 for details about this database).

For our experiments, the following distortion types were selected from the TID2013: quantization noise, mean shift (intensity shift), and change of color saturation. We selected this subset of distortions because they encompass the most important color related distortions in current imaging technologies for natural scene color images. For instance, quantization noise is closely related to color quantization. Intensity shift and change in color saturation are well-known distortions produced by color matching algorithms, color mapping algorithms and multiview imaging systems (Baranczuk et al., 2010; Wei and Mulligan, 2010; Fezza et al., 2014). The remaining 21 distortions from TID2013 database were not used in the experiments of this Chapter not even those affecting color because they incorporate also spatial distortions which typically impact the quality of the image much more strongly than CDs. Therefore, the human scores would be then more likely predominantly influenced by the spa-

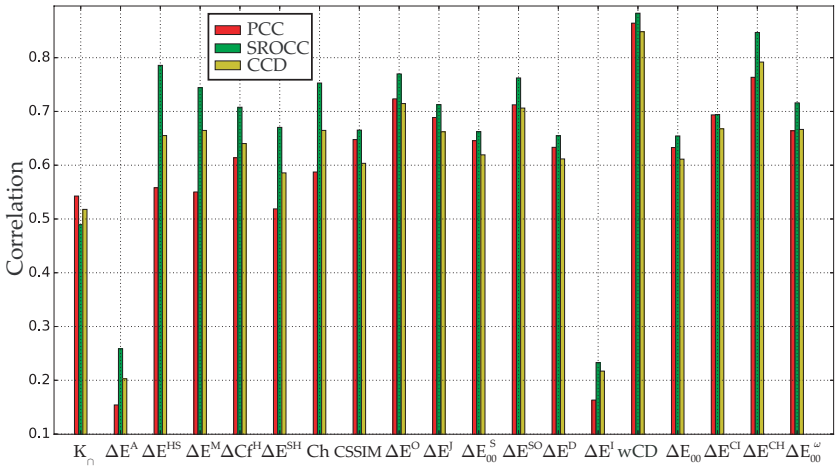

Figure 6.5: Performance of the considered 19 CD measures (18 existing and the proposed wCD) appraised on the color subset of TID2013 database. Performance is given in terms of the PCC, the SROCC and CCD between a given CD measure and the corresponding subjective scores.

tial distortions and not the color ones. For instance, we do not use chromatic aberrations and color quantization with dither because even though they have a large influence on color noise, they also produce strong artifacts of spatial nature such as blurring, false edges and/or rainbow edges which impact the "spatial" quality of the image much more strongly than its CD. In the case of contrast changes, we have shown in Chapter 4 that these image differences are better modeled by using the ratio of intensity values.

As we discussed in Chapter 4, the MOS values from TID2013 were collected using a methodology known in psychophysics as two alternative forced choice (2AFC) match to sample (Ponomarenko et al., 2015). In 2AFC three images are displayed (the reference and two distorted images) and an observer selects one of the two distorted images which they judge as more similar to the reference. That is, human observers are asked to select among two images the image that perceptually differs less from a reference (Kingdom and Prins, 2010b). Thus, the evaluation is made in terms of the presented current stimuli. Since the 2AFC was made within the "color" subset of the TID2013, the MOS scores designated to that subset are a measure of the color difference with respect to the reference image perceived by the observers. Therefore, TID2013 allows the individual analysis of certain distortion type or subset of distortion types (Ponomarenko et al., 2015).

Table 6.2: Percentage increase of the performance appraised on TID2013 of the proposed color difference measure (wCD) compared with the state-of-the-art methods.

	Percentage increase		
Measure symbol	**PCC**	**SROCC**	**CCD**
ΔE_{00}	52	72	67
$\Delta E_{00}^{\mathrm{S}}$	48	69	64
ΔE^{M}	84	41	48
$\Delta \mathrm{Cf}^{\mathrm{G}}$	59	53	56
CSSIM	47	68	70
Ch	69	38	48
$K_{\bigcap}$	87	153	107
ΔE_{00}^{ω}	42	50	47
ΔE^{I}	592	470	438
ΔE^{O}	19	26	24
ΔE^{SO}	25	33	31
ΔE^{J}	34	51	49
ΔE^{D}	52	72	67
ΔE^{HS}	80	27	51
ΔE^{A}	633	411	478
ΔE^{SH}	98	66	77
ΔE^{CI}	33	58	47
ΔE^{CH}	13	9	10

6.4.2 Overall performance of the tested measures

Figure 6.5 shows the PCC, the SROCC and the CCD appraised on the color subset of TID2013 database. The best performing CD measures from the state-of-the-art are ΔE^{CH}, ΔE^{O} and ΔE^{SO} displaying a strong correlation (correlation between the CD measures and the subjective scores higher than 0.7). However, note that the proposed image CD measure (wCD) outperforms those CD image measures (correlation between the proposed CD measure and the subjective scores higher than 0.8). Table 6.2 shows the percentage increase of the proposed method compared with the other state-of-the-art measures based on the correlation coefficients shown in Figure 6.5 after applying the Fisher's z transform. The percentage increase shows that the proposed methodology outperforms all other 18 image CD measures tested in this Chapter.

The worst performance across the three color distortion types is achieved by ΔE^{I}, ΔE^{A}, $K_{\bigcap}$ displaying a weak correlation (correlation between the CD measures and the subjective scores lower than 0.5). The poor performance of ΔE^{I} may be due to the fact that the measure focuses on complex spatial interactions such as perception of contrast, graininess, and sharpness while in fact it should focus on homogeneous textured areas (Deng et al., 1999). Although ΔE^{A} is an adaptive technique, the CD measure is computed using the RGB color space which is well-known to disagree with human perception of color. $K_{\bigcap}$ performs better but still the correlation is weak compared with the other tested methods.

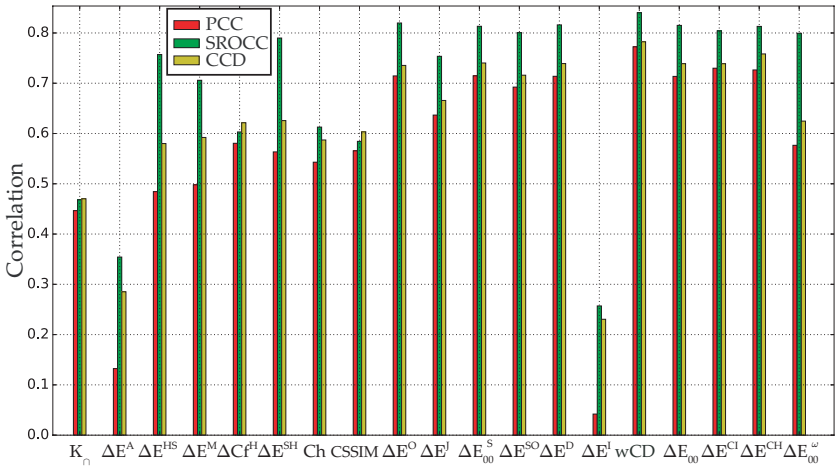

Figure 6.6: Performance of the considered CD measures appraised on TID2013 color saturation subset. Performance is given in terms of the PCC, the SROCC and CCD between a given CD measure and the corresponding subjective scores.

We also explore the performance of the tested CD measures on the individual distortion types to assess the strengths and weaknesses of the tested measures.

Figures 6.6, 6.7 and 6.8 show the PCC, SROCC and CCD appraised on TID2013 database per individual color distortion type, color saturation, mean shift and quantization noise, respectively. In the quantization noise the best performing is the Ch followed by ΔE^{J} and the proposed methodology wCD (Figure 6.8). The proposed methodology shows to be the best performing in the color saturation subset with a strong correlation (correlation between the proposed CD measure and the subjective scores higher than 0.8), see Figure 6.6. Also, wCD is one of the best performing methods together with ΔE^{CI} in the mean shift subset (Figure 6.7).

6.4.3 Discussion

Note that the good performance of Ch in the quantization noise subset is partially due to the fact that Ch compares the color distribution on the YCbCr color space (unlike any other of the considered methods) and TID2013 quantization noise was processed on the same color space. This suggests that color quantization noise can be evaluated by comparing the color distribution of the images when the comparison is made on the same operational color space where the distorted image was processed. Indeed, since color quantization modifies considerably the distribution of the color histogram in the given color space, a comparison of the distribution in the same space comes forward as an appropriate tool for this type of task. However, Ch performs poorly in the rest of the

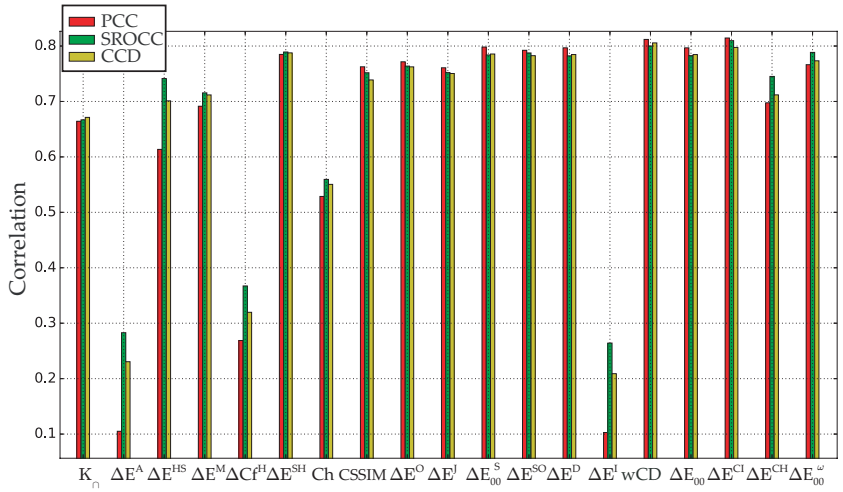

Figure 6.7: Performance of the considered CD measures appraised on TID2013 mean shift subset. Performance is given in terms of the PCC, the SROCC and CCD between a given CD measure and the corresponding subjective scores.

tested data because the other color related distortions (mean shift and change in color saturation) do not have a considerably impact in the color histogram of the images making Ch measure ineffective for this type of distortions.

Also note that there are no significant differences between ΔE_{00}, $\Delta E^{\mathrm{S}}_{00}$ and ΔE^{D}, i.e., there is a negligible improvement in terms of PCC, SROCC and CCD with subjective scores when a spatial filtering simulating the blur property of the human eyes effect is applied before computation of pixel-wise differences (cf. the spatial processing described by (Zhang and Wandell, 1997)). We attribute this behavior to the fact that CDs are perceived easier in large homogeneous areas where there is no contrast masking while CDs in small texture areas with color fluctuations are more difficult to perceive than in large homogeneous areas. Therefore, the spatial processing (band-pass filtering simulating blur property of human eyes as proposed by (Zhang and Wandell, 1997)) is an ineffective mechanism because the CD formulas are still applied pixel-wise instead of computing region based differences which is more appropriate due to the fact that humans perceive CDs easily in homogeneous textured areas. This is also confirmed by the results shown in Figures 6.5, 6.6, 6.7 and 6.8 where the proposed methodology (wCD) shows to be the best performing over all subsets of data.

The results show that overall, among all three considered sources of image color distortion, the best performing CD is the proposed methodology wCD displaying a strong correlation (correlation between the proposed CD measure and the subjective scores higher than 0.8) for all tested data. ΔE^{O}, ΔE^{SO}, ΔE^{M}, $\Delta \mathrm{Cf}^{\mathrm{H}}$, CSSIM, Ch, $K_{\bigcap}$, ΔE^{I} and ΔE^{HS} display a moderate correlation

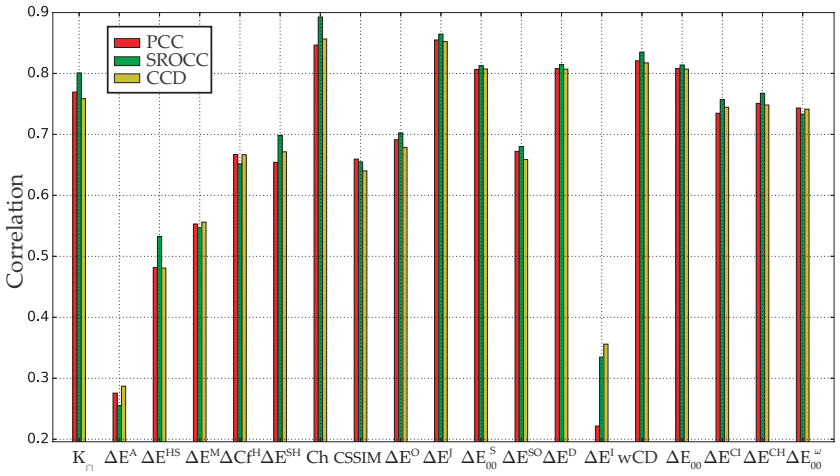

Figure 6.8: Performance of the considered CD measures appraised on TID2013 quantization noise subset. Performance is given in terms of the PCC, the SROCC and CCD between a given CD measure and the corresponding subjective scores.

(correlation between the CD measures and the subjective scores lower than 0.7) for all data. The worst performing methods are ΔE^{A} and ΔE^{I} displaying a weak correlation (correlation between the CD measures and the subjective scores lower than 0.5) for all tested data.

Revising individual color distortions, the previous experiments and results reveal that ΔE_{00}, $\Delta E_{00}^{\mathrm{S}}$, Ch, ΔE^{J}, ΔE^{D} and wCD are the best candidates to be used in color quantization application displaying a strong correlation with subjective scores in the color quantization subset. Also, the results show that the best candidates to assess images affected by black level shift are wCD and ΔE^{CI}. Additionally, the following CD measures are the best candidates for assessing CDs on images affected by change of color saturation: ΔE_{00}, $\Delta E_{00}^{\mathrm{S}}$, ΔE^{O}, ΔE^{CH}, ΔE^{D} and wCD displaying a strong correlation with subjective scores (SROCC).

Finally, note that the weights of the proposed methodology (α and β) in Equation (6.17) can be further investigated and tuned for different applications according to the importance of the differences in each color component.

6.5 Conclusions

This Chapter has reviewed and evaluated CD measures in natural scene color images. We tested eighteen state-of-the-art CD measures on selected data from one public database. To stimulate further experimentation, we made all the tested methods freely available as a plugin on the iFAS software tool. We selected our test image data such that the following applications are included:

color correction, color quantization, color mapping, color image similarity and retrieval. The images in these applications are typically affected by CDs due to quantization noise, intensity shift, contrast change, change in color saturation and change in color balance. Moreover, we have proposed a novel methodology for computing color difference in natural scene color images based on the findings of the state-of-the-art review; the proposed method is named wCD.

Our experiments show that ΔE_{00}, $\Delta E^{\mathrm{S}}_{00}$, ΔE^{D} and wCD achieve a strong correlation with subjective scores in the mean shift subset. In the quantization noise the best performing are the Ch followed by ΔE^{J} and the proposed methodology wCD. The following CD measures are the best candidates for assessing CDs on images affected by change of color saturation: ΔE_{00}, $\Delta E^{\mathrm{S}}_{00}$, ΔE^{O}, ΔE^{CH}, ΔE^{D} and wCD showing a strong correlation with subjective scores. Overall, the proposed methodology (wCD) is clearly the best performing CD measure tested in this work.

Additionally, we found that relying on descriptive statistics from pixel-wise differences is unreliable for computing color differences typically perceived and reported by human observers. The results suggest that there are not significant differences in terms of correlation with subjective scores between ΔE_{00}, $\Delta E^{\mathrm{S}}_{00}$ and ΔE^{D}. This is important because it indicates that many CD measures for images are designed using an ineffective mechanisms for computing CDs, i.e., the computation of pixel-wise differences after preprocessing based on filtering. However, it is well-known that humans perceive CD better in flat areas than in complex structures (Deng et al., 1999). Thus, it will be more desirable to measure CDs in homogeneous patches (based on image segmentation) and then combine them into an overall CD as the proposed methodology. This is confirmed as well by the good performance achieved by the proposed methodology which is based on computation of local differences in homogeneous textured patches.

The contributions reported in this Chapter resulted in one international conference proceedings (Ortiz-Jaramillo et al., 2016a), and one peer-reviewed journal paper (Ortiz-Jaramillo et al., 2018b).

7

Concluding remarks

7.1 Conclusions

In this thesis we have studied application-specific fidelity assessment models with the purpose of predicting visual/perceived differences between the test image and its corresponding reference (original/unaltered) image that typically a human subject (observer) would report. Particularly, this thesis has studied the following fidelity-related use cases: quality estimation of compressed video sequences, evaluation of contrast ratio changes in images, assessment of appearance changes in texture and evaluation of color differences in natural scene color images. We have contributed to each of these areas by reviewing the existing methodologies, proposing new numerical fidelity measures, and experimentally evaluating performance of the different measures.

This dissertation has studied quality evaluation of compressed video sequences in Chapter 3. We have evaluated and tested four of the most well-known state-of-the-art video quality measures on five different public video quality databases. Additionally, this thesis has proposed a methodology to advance existing video quality measures by introducing video content related indexes in their computation. The accuracy of the proposed method is comparable with the other state-of-the-art methods. Also, our experimental results have shown that unlike other conventional methods, the proposed method is of low complexity and satisfies the requirements of real-time applications. For example, in this thesis we have implemented a Python script able to compute perceived video quality at 12, 25 and 75 frames per second for 1920×1080, 1280×720 and 720×380 pixels, respectively. The main drawback of this proposed video quality measure is that an off-line training is needed with enough samples representing the wide range of quality levels, extent of details and motion.

In Chapter 4, we have investigated the problem of the assessment of contrast changes in images. We have performed an extensive experimental evaluation based on a total of six image contrast ratio measures, each evaluated and tested on two image quality assessment databases. We have proposed a novel methodology to compute contrast ratio in images by using local content

analysis. We have used Weber and Michelson contrast ratio formulas on small patches to simulate the cases where a small structure of interest is present on a uniform background or a square-wave grating of one cycle, respectively. The proposed method is able to predict changes (decrements, increments) in contrast more accurately than the other state-of-the-art algorithms. We have tested our methodology on a real case scenario: detection of changes in contrast level in interventional x-ray images acquired with varying dose. The results show that the proposed contrast ratio measure agrees with the subjective evaluation of interventionalists in interventional x-ray images.

In Chapter 5, this thesis has reviewed and evaluated fourteen texture analysis descriptors for automatic assessment of appearance changes in texture. Additionally, this thesis has discussed the impact of the parameter selection of the evaluated texture analysis descriptors. We have conducted an extensive experimental evaluation based on a total of three image databases. The results show that the signal processing methods are the best performing with a strong correlation with human evaluation in cut and loop pile surface constructions. Therefore, we believe that future work in the evaluation of appearance changes in texture should be developed by using signal processing methods because they provide the advantage of filter selection/design. The results also showed that the considered texture analysis techniques perform poorly in textile floor coverings with (shag) long pile construction.

In Chapter 6, this dissertation has studied the evaluation of perceived color differences in natural scene color images. We have reviewed and evaluated eighteen state-of-the-art color difference measures and we have discussed their performances. The measures have been tested on a total of twenty five different source images and three different color-related distortions. We have proposed a novel method to compute color differences in natural scene color images based on the findings of the review. We based our measure on the fact that humans assess color differences in natural scene color images by comparing sets of connected pixels or small patches. Those patches are typically characterized for being homogeneous or for possessing a unique texture pattern. The results have shown that the proposed method is able to predict color differences reported by human observers with higher accuracy (higher correlation levels) than the other state-of-the-art algorithms.

Finally, the majority of different methods studied in this thesis (existing as well as those proposed in the thesis) have been made available for the research community as an open source software toolset, named image Fidelity Assessment (iFAS). iFAS is detailed in Appendix B. iFAS provides the following basic image fidelity assessment tools: computation of fidelity measures on a single pair of images and/or in a full database, visualization of pixel-wise image differences and histogram of the image differences, scatter plots and correlation analysis between human scores and objective measures. The correlation analysis is performed following the most recent recommendations for the process of image fidelity assessment evaluation such as global correlation comparison, pairwise comparisons of correlations per reference, regression

analysis and model building. In this software toolset we have collected all the image fidelity methods tested in this thesis. That is, there are available a set of seven image contrast ratio measures, nineteen image color difference measures, fourteen texture analysis algorithms and five general purpose image quality measures. Since the existing methods are implemented by the author of this thesis (not by their original authors), it is necessary to perform in the future an evaluation of the methods included with iFAS to verify the same output as proposed by the original authors. The primary motivation for opening iFAS to the community is code sharing for image fidelity evaluation which is typically not available for many of the state-of-the-art algorithms and only a very limited subset of conventional fidelity indexes are easily accessible, e.g., the PSNR, CIEDE2000 (Luo et al., 2001), SSIM (Zhou et al., 2004).

7.2 Future work

Note that the evaluation methods tested in this thesis do not consider the variability of the reported human scores and assign uniform weights to all fidelity levels. Therefore, a study of more advanced statistical techniques for benchmarking image fidelity algorithms is proposed as future work with the purpose of validating the results presented in this thesis.

In the quality evaluation of compressed video sequences, the study of other distortion types (type of artifacts) remains as future research with the purpose of further validating the results presented in this work. Particularly, it is very important to consider the type of artifacts under study because the methodology can perform well only under specific conditions, e.g., a specific codec. Therefore, it may be necessary to consider a specific mapping function for different sources of distortion. In other words, the estimation of mapping function parameters should include not only content related features but also the distortion type. Additionally, since different spatial and temporal pooling on PSNR may lead to different results, the study of different pooling strategies remains as future work.

In the assessment of contrast changes in images, the main point for improvement of this proposed methodology is that the local content information is used in the entire image without discriminating if the patch is a structure of interest or just background with the purpose of keeping a low computational time. One possible approach for overcoming this issue would be to explore potential benefits of computing with a different formula the local contrast ratio values in the background patches (local patches where no bimodal distribution is found).

In the automatic assessment of appearance changes in texture, a study of other descriptors of texture for improving the process of measuring appearance retention in floor coverings remains as a future work. For instance, it is necessary to explore descriptors as hairiness, pilling, change of pattern definition, change in color, among others for improving the results presented in this work. Combinations of texture features such as MRF in the wavelet domain have

been commonly used in the texture analysis techniques. That motivates future investigation into the combination of signal processing based techniques and model based approaches.

In the assessment of color differences in natural scene color images, an interesting topic for future work is to select better weights of the proposed methodology for different applications. Additionally, other statistics and color spaces can be considered in the detected homogeneous textured patches. As an example, according to the results presented in this thesis the Image CD based on circular hue (Lee and Rogers, 2014) is a good candidate for improving the proposed methodology. Future work should further extend the scope of evaluation by including additional publicly available image databases as well as other color related types of distortion (e.g. gamut mapping) with the purpose of validating the results and generalizing the findings of our work. Also, since there is a considerable increase of computer-generated image content (Gu et al., 2018), the evaluation of the proposed methodology in computer-generated images is proposed as future work.

Future research should also consider the combination of multiple image fidelity measures to unify different methods for image fidelity assessment for any given application. For instance, it would be desirable to build a multi-factor approach for image fidelity assessment by combining different changes on image characteristics, e.g., contrast changes, color differences, appearance changes in texture. This unification can be performed by using modern machine learning techniques such as support vector machines (Yang et al., 2017), neural networks (Lukin et al., 2015), convolutional neural networks (Kim and Lee, 2017), deep neural network (Gu et al., 2014; Lv et al., 2015), among others.

Appendices

A

Databases

In this Chapter we describe the used databases in this thesis. Note that, from the following database descriptions, we show that they are designed for measuring perceived image differences. Therefore, the databases are appropriated to test the performance of the measures explored in this work.

A.1 Tampere Image Database (TID2013)

TID2013 (Ponomarenko et al., 2015) is a database intended for evaluating image fidelity measures. This database contains 25 reference images and 3000 distorted images (25 reference images $\times$ 24 types of distortions $\times$ 5 levels of distortions). Source images (displayed in Figure A.1) are obtained from the Kodak Lossless True Color Image Suite (Kodak, 2013).

The distortions available in TID2013 are (distortions marked in bold produce changes in color) (Ponomarenko et al., 2015):

- additive Gaussian noise,
- **additive noise in color components**,
- **spatially correlated noise**,
- masked noise,
- high frequency noise,
- **impulse noise**,
- **quantization noise**,
- Gaussian blur,
- image denoising,
- **JPEG compression**,
- JPEG2000 compression,

Figure A.1: The 30 reference images used in TID2013 database.

- JPEG transmission errors,
- JPEG2000 transmission errors,
- non eccentricity pattern noise,
- local block-wise distortions of different intensity,
- **mean shift (intensity shift)**,
- **contrast change**,
- **change of color saturation**,
- multiplicative Gaussian noise,
- comfort noise,
- lossy compression of noisy images,
- **image color quantization with dither**,
- **chromatic aberrations**,
- sparse sampling and reconstruction.

Each distorted image has a subjective score for comparing the performance between fidelity measures. The subjective scores are expressed in terms of Mean Opinion Scores (MOS). The MOS values from TID2013 were collected using a methodology known in psychophysics as two alternative forced choice (2AFC) match to sample (Ponomarenko et al., 2015). In 2AFC three images are displayed and an observer selects a better image between two distorted ones. That is, the human observers are asked to select among two images the image that differs less from a reference (Kingdom and Prins, 2010b). Thus, the evaluation is made in terms of the presented current stimuli. Since the 2AFC was made in the TID2013, the MOS scores designated to the images are a measure of the image differences with respect to the reference image perceived by the human observers, i.e., the fidelity of the distorted images with respect to the reference. Therefore, TID2013 allows the individual analysis of certain distortion type or subset of distortion types (Ponomarenko et al., 2015).

A.2 Computational and Subjective Image Quality database (CSIQ)

The CSIQ database consists of 30 original images (see Figure A.2) and 720 distorted images (6 different types of distortions at 4 different levels of distortion). The reference images were obtained from the U.S. National Park Service. The CSIQ database includes the following distortion types (distortions marked in bold produce changes in color) (Larson and Chandler, 2010):

- **JPEG compression**,
- JPEG2000 compression,
- **global contrast decrements**,
- additive pink Gaussian noise, and
- Gaussian blurring.

Each distorted image has a subjective score for comparing the performance between fidelity measures. The subjective scores are expressed in terms of Differential Mean Opinion Scores (DMOS). The DMOS values from CSIQ database were collected based on a linear displacement of the images. That is, all of the distorted versions of an original image were viewed simultaneously on a monitor array and placed in relation to one another according to the perceived quality difference (Larson and Chandler, 2010). The images were sorted by the observers according to the perceived differences with respect to the reference. Thus, the DMOS scores designated to the distorted images are a measure of the perceived image differences with respect to the reference reported by the human observers, i.e., the fidelity of the distorted images with respect to the reference image.

Figure A.2: The 30 reference images used in CSIQ database.

A.3 Anthropomorphic chest phantom

We use a static anthropomorphic chest phantom scanned with and without a 10 cm polymethyl methacrylate plate to simulate standard and large chest thickness, respectively, at six dose levels. The static anthropomorphic chest phantom contains contrast filled coronary arteries (Radiology Support Devices Alderson Phantoms, Long Beach, USA) was scanned on an Allura interventional X-ray system (Philips Healthcare, Best, The Netherlands). The dose levels and subjective scores of the two static anthropomorphic chest phantoms are shown in Table A.1. The images in Figure A.3 were evaluated by four interventionalists (cardiologist/radiologist) from Ghent University Hospital resulting in a mean opinion score per image (Kumcu et al., 2015b). The interventionalists rated the the similarity of each pair of images using a continuous scale from 0 (completely different) to 100 (exactly the same). See (Kumcu et al., 2015b) for further details about this database.

Table A.1: Table of dose levels and subjective scores of the two static anthropomorphic chest phantoms.

Dose [μGy/s]	2613	1973	1302	978	643	308
MOS	77	73	72	74	52	33
Dose [μGy/s]	9330	6700	4980	4140	3320	2475
MOS	75	69	66	63	52	46

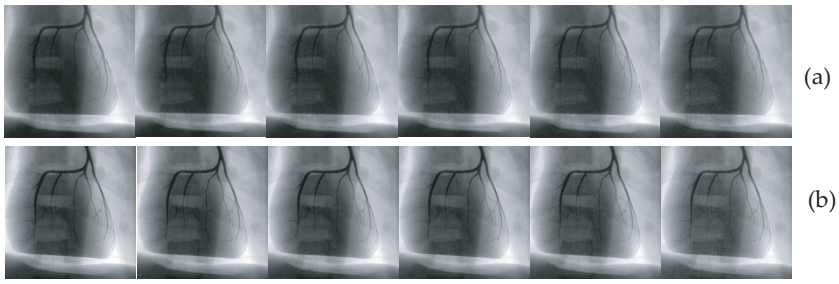

Figure A.3: Anthropomorphic chest phantom scanned (a) with [from left to right chest phantom with 10 cm polymethyl methacrylate scanned at 9330 μGy/s, 6700 μGy/s, 4980 μGy/s, 4140 μGy/s, 3320 μGy/s, 2475 μGy/s:] and (b) without [from left to right chest phantom scanned at 2613 μGy/s, 1973 μGy/s, 1302 μGy/s, 978 μGy/s, 643 μGy/s, 308 μGy/s,] a 10 cm polymethyl methacrylate.

A.4 Carpet reference standards

The basic material used in fabrics is a fibre either natural or synthetic. Thus, fibres are well know and used for carpet manufacturing. Usually, this fibres are twisted together or only grouped together to made yarns. Often two or more yarns are entwined to make a thicker yarn allowing manufacturers to obtain technical or aesthetic special effects. With the purpose of constructing carpets, tuft of yarns are stitched into a woven or non woven fabric to form the face material (this process is called tufting). Thanks to the tufting process carpets are available in increasingly variety of styles, colors and aesthetics. This variety of carpets is possible by changing the surface construction. According to (ASTM-D5684-10, 2010; ISO-2424:2007, 2007), there are three basic surface constructions of tufted carpets (loop, cut and cut/loop pile).

The surface construction types are divided as follows:

- **Cut pile** is a type of carpet that involves a cut of the loops that are created during the weaving process (See Figure A.4).
  - *Velours* is a type of cut pile carpet with a close pile density giving a very flat surface.
  - *Saxony* refers to a cut pile carpet from which yarns are a little bit longer and thicker than Velours. As consequence it is possible to

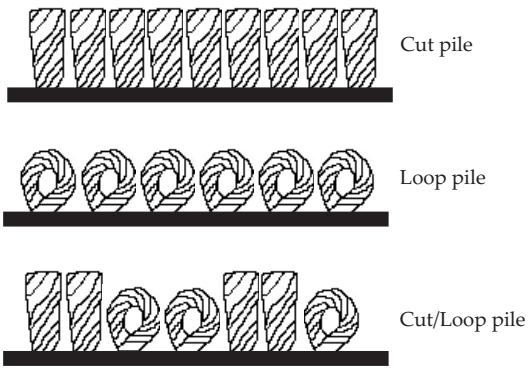

Figure A.4: Standard surface construction of tufted carpets.

differentiate better the point effect of the yarn.

- *Frisé* is a type of cut pile carpet which is made from high twisted fibers. After cut its yarns, the piles take different directions.
- *Shag* is a carpet with twisted yarns that have a special pile height (about 2.5cm or higher). This high distance between the face material and the tufts causes that the pile fall down.

- **Loop pile** is a design of carpet that it is created by using a series of uncut loops (See Figure A.4).
  - *Level loop* is a type of carpet that uses loops of the same size; creating a smooth surface.
  - *High/low or patterned loop* is a carpet that offers slightly different variations in loop height creating a pattern in the carpet.
  - *Textured loop* is a type of carpet that exhibits some pile height variation. Although the pile height differentiation is usually slight and has little or no pattern definition.

- **Cut/loop pile** is created by tufting some loops higher than others. When the carpet is sheared, the higher loop tufts are cut but the lower ones are not (See Figure A.4). The resulting cut pile tufts looks darker than the loops, creating a pattern which forms interesting designs.
  - *Cut and loop* is a high/low construction where patterning is achieved by *high pile* (cut pile) and *low pile* (loop pile).
  - *Tip sheared loop* is made from two sets of loops of different heights. After production, the higher loops are sheared to provide a soft handle appearance.
  - *Level-sheared carpet* is obtained by shearing off the protruding loops to the same height as the non-cut pile yarns.

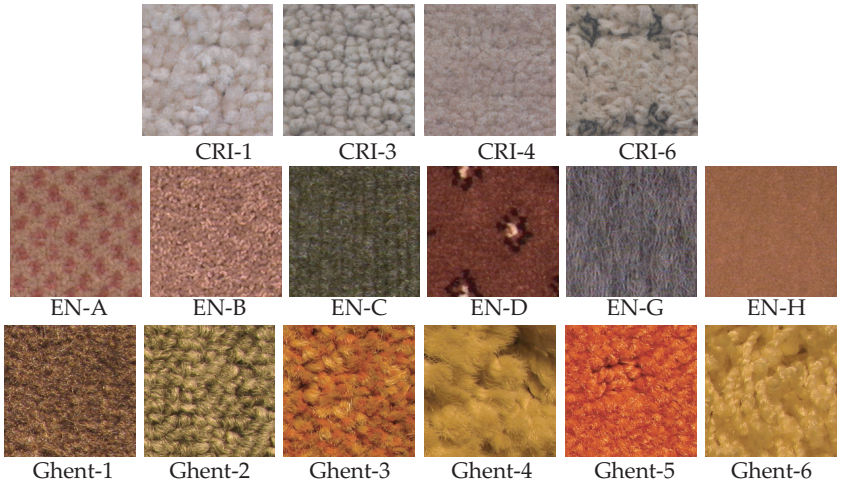

Figure A.5: The 16 reference sets of textile floor coverings used in this thesis.

– *Level cut/loop* is made by weaving even loops of yarn into carpet backing at both ends.

The used reference sets of textile floor coverings are shown in Figure A.5. The Figure shows textile floor coverings with construction types using level loop (CRI-3, EN-A and Ghent-2), cut Saxony (CRI-1, CRI-4, EN-B, EN-C, EN-D and Ghent-3), tip-sheared loop (CRI-6), cut/frisé (EN-G), woven velours (EN-H and Ghent-1), cut pile shag (Ghent-4 and Ghent-6) and cut/loop pile (Ghent-5). For more information about characteristics of carpet types and construction types see for example (ISO-2424:2007, 2007; ASTM-D5684-10, 2010; Orjuela-Vargas, 2012).

A.5 Video quality databases

The video quality databases used in this thesis are composed of video quality sequences compressed using H.264 codec (Pechard et al., 2011; Zhang et al., 2011; Pitrey et al., 2012). Note that the `CIF` and `4CIF EPFL-PoliMI Video Quality Assessment Database` (De-Simone et al., 2009) have videos compressed with H.264 followed by packet loss simulation. Particularly, the following databases are used to test the state-of-the-art numerical video quality measures: the `IRCCyN IVC 1080i: an HD video quality database` (Pechard et al., 2011), the `IVP Subjective Video Quality Database` (only the compressed sequences using H.264 codec) (Zhang et al., 2011), the `IRCCyN IVC Influence Content` (Pitrey et al., 2012), the `CIF EPFL-PoliMI Video Quality Assessment Database` (De-Simone et al., 2009) and the `4CIF EPFL-PoliMI Video Quality Assessment Database` (De-Simone et al., 2009). The majority

Figure A.6: The 20 reference sequences from the `IRCCyN IVC 1080i` database.

of the databases tested in this thesis are evaluated by using single-stimulus and the absolute category rating scoring system (using a [1-5] scale). The `CIF` and `4CIF EPFL-PoliMI Video Quality Assessment Databases` (De-Simone et al., 2009) uses single-stimulus and the continues category rating scoring system (using a [1-5] scale). The distorted sequences were subjectively evaluated for at least 25 human subjects.

The following is the list of source sequences from the tested public video quality databases:

- The `IRCCyN IVC 1080i: an HD video quality database` (Pechard et al., 2011) contains 20 source video sequences of resolution 1920×1080 at 25 frames per second (see Figure A.6): (1) *above marathon*, (2) *captain*, (3) *concert*, (4) *credits*, (5) *dance in the woods*, (6) *duck fly*, (7) *foot*, (8) *fountain man*, (9) *golf*, (10) *group disorder*, (11) *inside marathon*, (12) *movie*, (13) *new parkrun*, (14) *rendezvous*, (15) *show*, (16) *standing*, (17) *stockholm travel*, (18) *tree pan*, (19) *ulriksdals*, and (20) *voile*.

- The `IVP Subjective Video Quality Database` (Zhang et al., 2011) contains 10 source video sequences of resolution 1920×1088 at 25 frames per second (see Figure A.7): (21) *bus*, (22) *laser*, (23) *overbridge*, (24) *robot*, (25) *shelf*, (26) *square*, (27) *toys calendar*, (28) *tractor*, (29) *train*, (30) *tube*.

- The `IRCCyN IVC Influence Content` (Pitrey et al., 2012) contains 60 source video sequences (see Figure A.8) of resolution 960×540 at 25 frames per second (this database is used only for testing, i.e., none of its samples were used during training, inspection and/or selection of the *content related indexes* and/or mapping functions): (31) *animation 1*, (32)

Figure A.7: The 10 reference sequences from the `IVP` database.

space shuttle, (33) *kitesurfing 1*, (34) *ducks*, (35) *station*, (36) *kitesurfing 2*, (37) *factory 1*, (38) *skateboarding 1*, (39) *crew*, (40) *intotree*, (41) *touchdown*, (42) *kitesurfing 3*, (43) *aspen*, (44) *pedestrian area*, (45) *skateboarding 2*, (46) *city*, (47) *night traffic*, (48) *mother in the woods*, (49) *skateboarding 3*, (50) *fire*, (51) *red kayak*, (52) *day traffic*, (53) *dinner 1*, (54) *west wind easy*, (55) *rush hour*, (56) *big buck bunny*, (57) *kitesurfing 4*, (58) *RC cars*, (59) *old town cross*, (60) *hiking 1*, (61) *RC*, (62) *hiking 2*, (63) *halftime show 1*, (64) *bee*, (65) *boxing*, (66) *teaching*, (67) *halftime show 2*, (68) *cruise*, (69) *animation 2*, (70) *river bed*, (71) *life*, (72) *christmas*, (73) *waterfall*, (74) *dinner 2*, (75) *factory 2*, (76) *dinner 3*, (77) *tractor*, (78) *rush field cuts*, (79) *mobile*, (80) *excavator*, (81) *basketball*, (82) *sitting on the beach*, (83) *walking on the beach*, (84) *credits 1*, (85) *bridge*, (86) *crowd wave*, (87) *park joy*, (88) *crowd run*, (89) *credits 2*, and (90) *parade*.

- The `CIF EPFL-PoliMI Video Quality Assessment Database` (De-Simone et al., 2009) contains 6 source video sequences (see Figure A.9) of resolution 352×288 at 30 frames per second (this database is used only for testing, i.e., none of its samples were used during training, inspection and/or selection of the *content related indexes* and/or mapping functions): (91) *foreman*, (92) *hall*, (93) *mobile*, (94) *mother*, (95) *news*, and (96) *paris*.

- The `4CIF EPFL-PoliMI Video Quality Assessment Database` (De-Simone et al., 2009) contains 6 source video sequences (see Figure A.9) of resolution 704×576 at 25 frames per second (this database is used only for testing, i.e., none of its samples were used during training, inspection and/or selection of the *content related indexes* and/or mapping functions): (97) *crowdrun*, (98) *duckstakeoff*, (99) *harbour*, (100) *ice*, (101) *parkjoy*, and (102) *soccer*.

Figure A.10 shows the scatter plot of SA and TA for the used databases. SA and TA are the mean value of the magnitude of the SI13 image and the mean total variation over all frames of the temporal gradient (cf. s_1 and t_1 in Section 3.3.2), respectively. The scatter plot shows that the variety of spatial and temporal activity levels in the video test sequences is high, i.e., a wide

range of extent of image details and motion. The plot together with previous database descriptions show that a wide range of video content is used. They range from very low motion (news) to very high motion (sports) and from low textured (cartoons) to high textured (natural scenes) sequences.

Figure A.8: The 60 reference sequences from the `IRCCyN IVC Influence Content` database.

Figure A.9: The 12 reference sequences from the `CIF` and `4CIF` `EPFL-PoliMI Video Quality Assessment IVC` database.

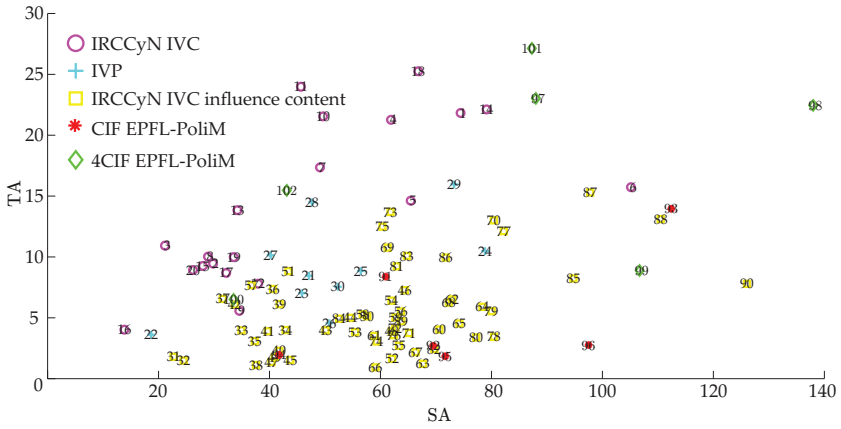

Figure A.10: Scatter plot of SA and TA computed on all databases. SA and TA are the mean value of the magnitude of the SI13 image and the mean total variation over all frames of the temporal gradient (cf. s_1 and t_1 in Section 3.3.2). Labels indicate the source sequence.

B

Image Fidelity Assessment software

image Fidelity Assessment (iFAS) is a software tool designed to assist image quality researchers providing easy access to a range of state-of-the-art measures which can be applied on a single pair of images and/or in a full database, as well as intuitive visualizations that aid data analysis, e.g., images and histograms of pixel-wise image differences, scatter plots and correlation analysis. The software is freely available for non-commercial use.

In this Chapter we describe iFAS software tool proposed in this thesis for assisting researchers, engineers and other users in the process of image fidelity assessment.

Figure B.1 shows a screenshot of iFAS user interface. The interface displays seven main components: (1) the main menu, (2) the reference image, (3) the test image, (4) the pixel-wise difference image, (5) the histogram of the pixel-wise difference image, (6) the scatter plot of the MOS versus the measure values for a given reference, and (7) the results pane. iFAS includes the following types of analysis:

- Single Source - Single Sample: this type of analysis computes the selected fidelity measures between one reference image and one corresponding test image, i.e., the current images displayed on the user interface (images are selected using a file chooser).

- Single Source - Multiple Sample: this type of analysis computes the fidelity measures selected by the user between one reference image and a number of corresponding test images as specified by the user using a file chooser.

- Multiple Source - Multiple Sample: this type of analysis computes the fidelity measures selected by the user between a number of reference images and a number of test images. This is like executing a Single Source - Multiple Sample several times using different reference images and their corresponding test samples. Note that, the user has the responsibility

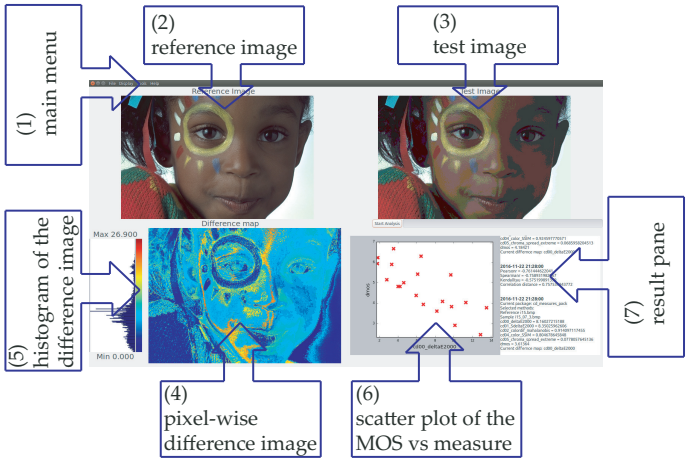

Figure B.1: iFAS main window screenshot.

to make sure that the test/distorted images correspond in scene to the reference/source images.

iFAS includes a detailed help document (included with iFAS files) with all the instructions for the recommended use of iFAS for maximum utility. Also, every analysis can be saved and loaded for further analysis avoiding unnecessary multiple computations of the same data. iFAS also allows visualization of scatter plots between two fidelity measures or a fidelity measure and subjective scores for a given source image. That is, it is possible to visualize the relationship between measures for a specific image content.

Additionally, iFAS computes and displays in the results pane the following correlation indexes typically used as performance indicators of fidelity measures: PCC (Chen and Popovich, 2002a), SROCC (Chen and Popovich, 2002b), KRCC (Chen and Popovich, 2002c) and CCD (Székely et al., 2007). Note that these indexes are computed with the values currently displayed on the scatter plot component (only for Single Source - Multiple Sample and Multiple Source - Multiple Sample).

In general, many fidelity measures pre-compute pixel-wise differences and later compute a global fidelity measure from descriptive statistics, termed spatial pooling. Therefore, iFAS also displays this intermediate result with the purpose of providing local information about the fidelity measure behavior. In the cases where the fidelity measure is based on features, iFAS computes those features on local windows and then displays the difference of the local computations (pixel-wise image difference). The pixel-wise image difference has associated a 255 bin histogram. This histogram can be considered to select the appropriated statistics for computing the global fidelity index during a fidelity measure design process.

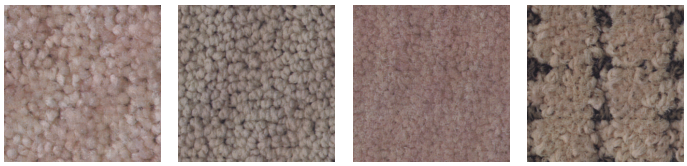

Figure B.2: Cutouts of textures evaluated from the CRI standard.

In the alpha version, we include 18 color difference measures for digital images (see Chapter 6), 14 texture analysis algorithms (see Chapter 5), 6 image contrast ratio measures (see Chapter 4) and six image quality measures for objective fidelity assessment.

B.1 How do I use iFAS?

In this Section we describe and explore two use case scenarios with the purpose of showing the advanced functionality of iFAS for the analysis of objective fidelity measures. The use cases are the benchmarking of 6 texture extraction algorithms on a textile floor covering database and the evaluation of 3 color correction algorithms for multiview imaging using 2 image color difference measures. The set of methods used for the analysis in this section were selected from (Ortiz-Jaramillo et al., 2014b, 2016a).

B.1.1 Application 1: evaluating appearance changes in textiles

iFAS can help in the evaluation of appearance changes in textiles in two ways. First in the benchmarking of possible candidate methods for measuring fine changes of global texture. The benchmarking consists of global correlation comparison, correlation analysis per reference sample. And second by finding an appropriated model using regression analysis for measuring the wear degree of future samples.

As an use case, we use the set of images composed of scanned printed images from the CRI standard photo set (CRI Test Method 103, 2015). The set of references (see Figure B.2) include texture types with loop (CRI-3), cut (CRI-1 and CRI-4) and tip-sheared (CRI-6) (see Appendix A.4 for more details about carpet construction). We use 32 different reference samples extracted using random cutouts (8 per CRI reference). For instance, CRI-1 has 8 cutouts representing the reference samples. Similarly, for each sample belonging to a specific wear label, we extract 2 cutouts, i.e., 16 test samples per reference. That is a total of 512 test [(16 cutout test samples per reference) $\times$ (32 reference samples)] and 32 reference samples (8 cutouts $\times$ 4 CRI samples).

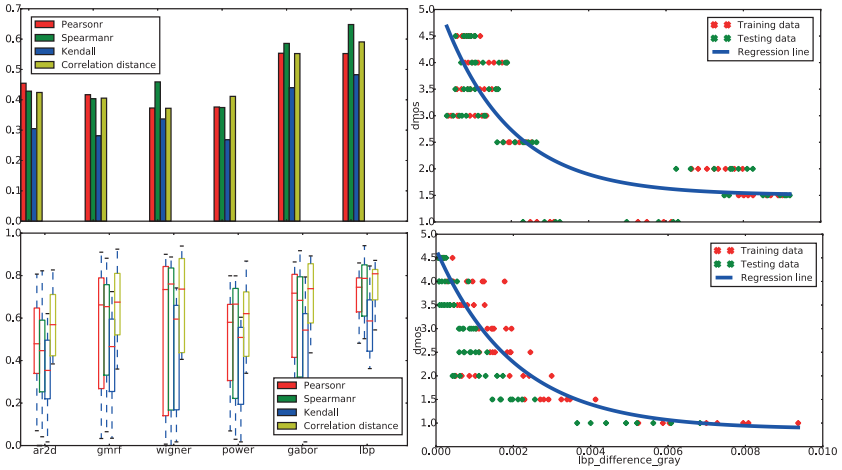

Figure B.3: Image saved in eps format using iFAS. (top left) bar plot global correlation, (bottom left) box plot of the performance per individual reference, (top right) model for CRI-3 and (bottom right) model for CRI-6

Here, we benchmark 6 different texture analysis algorithms (see Chapter 5 for details on the algorithms and the parameter selection): power spectrum (PS) (Weszka et al., 1976), Gabor filtering (GF) (Manjunath and Ma, 1996), Pseudo Wigner distribution (PWD) (Cristobal and Hormigo, 1999), autoregressive (AR) models (Joshi et al., 2009), Local Binary Patterns (LBP) (Maenpaa, 2003) and Gaussian Markov Random Fields (GMRF).

After computing the texture difference between samples using iFAS Multiple Source - Multiple Sample analysis tool, the global correlation is computed between the obtained data and the wear labels included with the CRI database. The global correlation analysis is very useful to identify the performance of the tested methods across different content, in this case CRI standard type, i.e., it helps to identify if only one model is necessary to perform the task. This performance comparison is also performed by iFAS software by using the Global correlation analysis tool.

The Figure B.3 (top left) shows the figure saved in eps format using iFAS software corresponding to the bar plot for global correlation analysis. The software also shows on the result pane the correlation values and the best performing methods according those values, e.g.,

```
Best according to PCC is gabor_features_difference:  0.55339

Best according to SROCC is lbp_difference_gray:  0.64797

Best according to KRCC is lbp_difference_gray:  0.48265

Best according to CCD is lbp_difference:  0.48265
```

These are not very high correlations to build a model. However, since the texture patterns for the carpets are standardized on the textile industry, it is also valid to build a model per each CRI standard type. Therefore, box plot analysis is very useful for identifying how well the measures perform if the content remains the same (in this case the CRI standard type). iFAS also possess this kind of analysis. Figure B.3 (bottom left) shows the box plot for the considered texture features per reference sample resulting from iFAS. Additionally, the software provides on the result pane a pairwise comparison between the methods with the purpose of identifying if there are statistically significant differences in terms of the performance between the tested methods. Also, the software presents the best performing methods according to the pairwise comparisons (here only the example for CCD):

```
Best method from pairwise comparison between Correlation
distance

lbp and it is statistically significant better than:

ar2d p_value:  0.00000

gmrf p_value:  0.00014

power p_value:  0.00000
```

This indicates that LBP technique is the best performing method and it is statistically better than AR, GMRF and PS. Also, LBP performs equally or better than GF and PWD. Therefore, further analysis is performed using LBP, i.e., the modeling of each individual CRI standard type.

After selecting the best performing method, it is very useful to build a model with the current data to be used on new samples. In iFAS, we simulate that process by building models using the current samples. The user needs to select among the samples the training and test subsets. Here we show on Figure B.3 (top-bottom right) the regression analysis for two CRI standard types, CRI-3 and CRI-6, respectively. The dots are the actual data and the line shows the fitted model using the training data. The parameters of the model are shown in the result pane together with the correlation coefficients for the model. This information comes handy for estimating the performance of the model for external use of the built models. We show as example the information displayed on the result pane for the CRI-6 type:

```
The optimal parameters for function
a0 + exp(a1 * x + a2)
a0:  0.87016
a1:  -494.71609
a2:  1.33914
Training values:
PCC: 0.88335
SROCC: 0.87563
CCD: 0.85311
Testing values:
PCC: 0.89326
SROCC: 0.90553
CCD: 0.88545
```

In summary, the methods presented in this Section illustrate how using iFAS in a very simple fashion could support a very complicated process such as benchmarking. Additionally, the software provides enough information and an internal decision making systems to suggest the best performing methods according to each analysis, e.g., the pairwise comparisons and the global correlation analysis.

B.1.2 Application 2: evaluating color correction in mutliview imaging

In multiview imaging, color correction is used to eliminate color inconsistencies between views. Here, the color correction algorithm modifies the color of one of the views ("color-processed" image) such that the color is visually equivalent to the other view ("reference" image). To select this algorithm, the correlation analysis does not provide the type of information we are looking for (the best color correction algorithm). Therefore, iFAS can help us instead by plotting in boxes and analyzing the color difference values between the images with the purpose of examining if there are color corrections with lower color differences. iFAS performs the analysis based on pairwise comparisons and on the assumption that "higher is better." Therefore, iFAS transforms to similarities (s) from the original differences (d) with the formula $s = 1 - d/\max_d$. Here, $\max_d$ is the maximum over the whole set of computed/available differences.

As an use case, we use a database containing 144 images: 3 references $\times$ 4 distortion types $\times$ 4 distortion levels $\times$ 3 color correction algorithms. The 3 reference images were selected from SRTMI database (Yeganeh and Wang, 2013). These images are color manipulated with the purpose of simulating the multiview environment using the following types of color disparities: change of the color contrast on one of the color channels and black level shift (4 different distortion levels for each distortion type). Afterwards the following color correction algorithms are independently applied on each color processed image: color correction based on transferring statistical properties (CCstats) (Reinhard et al., 2001) on CIELab color appearance model (CIE, 1976), color correction based on point correspondences (CCpoint) (Ly et al., 2014) and color correction based on region correspondences (CCregion) (Ly et al., 2015).

The 144 images are compared in terms of color by using the following color difference measures: the CIEDE2000 formula (Luo et al., 2001) and the just noticeable CD measure (Chou and Liu, 2007). These measures and all the measures study in Chapter 6 are available in iFAS software tool.

iFAS computes the color differences in the database and performs the pairwise comparisons using the obtained differences per source content (reference sample). Figures B.4 (top left-right and bottom left) show the result from such an analysis performed by iFAS. Also, iFAS provides the best performing methods according to the pairwise comparisons (we show as example the information displayed on the result pane for one of the references comparing the 3 color correction algorithms):

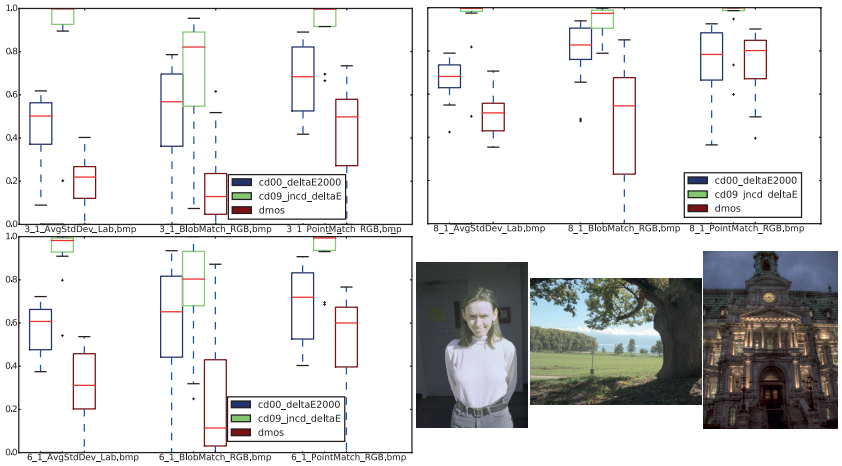

Figure B.4: Image saved in eps format using iFAS. (top left) bar plot color differences for SRTMI 3, (bottom left) bar plot color differences for SRTMI 6, (top right) bar plot color differences for SRTMI 8, and (bottom right) the three SRTMI reference samples from left to right SRTMI 3, 6 and 8, respectively.

```
Best method from pairwise comparison between cd00_deltaE2000

8_1_PointMatch_RGB.bmp and it is statistically better than:

Best method from pairwise comparison between
cd09_jncd_deltaE

8_1_PointMatch_RGB.bmp and it is statistically better than:

8_1_BlobMatch_RGB.bmp p_value:  0.01333

Best method from pairwise comparison between dmos

8_1_PointMatch_RGB.bmp and it is statistically better than:

8_1_AvgStdDev_Lab.bmp p_value:  0.00146

8_1_BlobMatch_RGB.bmp p_value:  0.00801
```

That is, according to the subjective scores (in the result pane named `dmos`) the best performing color correction algorithm is the method CCpoint, named `PointMatch_RGB` on the result pane. Note that the CIEDE2000 formula does not find any statistical significant differences and the just noticeable CD measure finds only statistical differences between CCpoint and CCregion but not between CCpoint and CCstats. However, humans find differences between CCpoint and CCstats as well as between CCpoint and CCregion. Therefore, the best performing color correction algorithm is CCpoint because it produces the

minimum perceived differences according to the subjective scores and the just noticeable CD measure. To finally decide, it is possible to refine the selection by looking at the reference, test and difference images displayed in the main window of iFAS.

C

Real-time estimation of perceived video quality

In this thesis we have implemented an algorithm for low complexity, real-time and content-aware estimation of perceived video quality. The target audience for this algorithm is described in the following paragraph.

- Video content providers and distributors (digital TV, video streaming applications):
  - to tune (de)coder and/or transmission parameters for their content for a target video quality at the end-user device. Adequately selected transmission parameters can bring up to 34% energy savings in a streaming system (around 0.05% of global energy consumption) (Ejembi and Bhatti, 2015);
  - to monitor video quality at the end user device, e.g., for compliance reporting and quality control.
- Video camera solution vendors as well as display technology vendors:
  - during the development cycle, to shorten the process, to optimize the design for desired/optimal performance of the products by using CPSNR to objectively guide design decisions, e.g., (de)coder settings, system parameter settings for specific environment conditions.
  - at runtime, as a building block of the control loop to adapt in real-time the camera/display system parameters such as (de)coder settings, based on the estimated video quality at a given moment and under the given conditions (video content, individual user preferences, among others).

Currently, the CPSNR algorithm has been successfully tested on publicly available benchmark databases (see Chapter 3). A non-optimized, yet already real-time, Python implementation is available. CPSNR is competitive in performance with conventional methods (correlation higher than 80% with human-perceived video quality trained/tested on public benchmarking databases) but

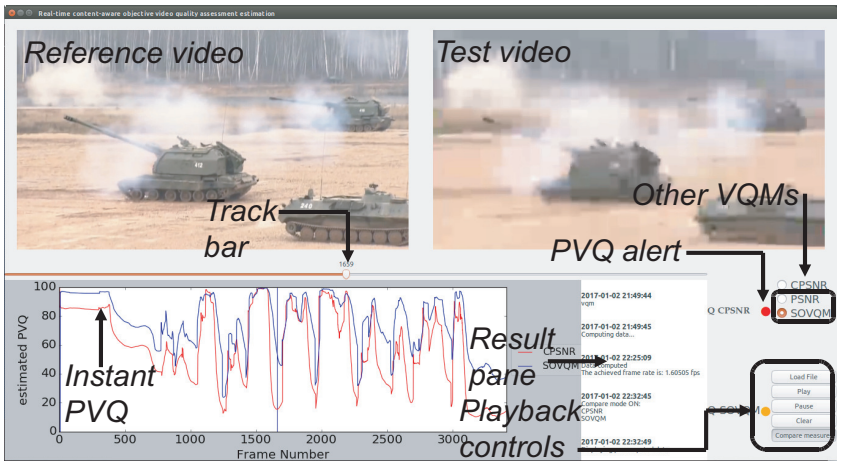

Figure C.1: Screenshot of the computer interface of the CPSNR implementation for video quality reporting and the control panel description of the interface.

it is of notably lower complexity (e.g., in comparison to the standardized method for objectively measuring video quality (SOVQM) (Pinson and Wolf, 2004a)). CPSNR runs at 12, 25 and 75 fps for 1920×1080, 1280×720 and 720×380 pixels, respectively, while the SOVQM runs at 1 fps for 720×380 pixels. The test data so far comprise the following:

- distortion introduced by H.264 compression and by packet loss;
- resolutions of 1920×1080, 960×540, 704×576 and 352×288 pixels;
- frame rates of 25 and 30 fps.

Unlike other conventional methods, CPSNR is of low complexity and satisfies the requirements of real-time applications. At the same time, accuracy of the CPSNR predictions is comparable with the state-of-the-art methods. The current implementation is in Python 2.7 (Python, 2016) using third party libraries (NumPy (SciPy, 2016), Python bindings to OpenCV (OpenCV, 2016)). No GPU or CPU acceleration is used yet, but it would be easy to do so.

Figure C.1 shows the screenshot of the video quality reporting tool implemented to demonstrate the power of CPSNR. This implementation has been successfully demonstrated in the Imec Technology Forum 2017 (Ortiz-Jaramillo, 2017a).

D

Isodata algorithm

In this chapter we describe the technique used in Chapter 4 for the computation of the mid point of the histogram and thus select a set of pixels likely to be inside the foreground (edge pixels) and another set likely to be in the background.

The `ISODATA` algorithm is an iterative technique for threshold selection. It finds the minimum between two modes of a histogram. The algorithm finds the threshold τ as follows:

1. Compute initial threshold

$$\tau_{\text{prev}} = \frac{1}{MN} \sum_{i,j} I(i,j),$$

   where $I(i,j)$, M and N are the (i,j)th pixel value, the number of rows and the number of columns of an image patch, respectively.

2. Divide the image patch into two sets of the current foreground and background (Ω^{F} and Ω^{B}, respectively), where $I(i,j) \in \Omega^{\text{F}}$ if $I(i,j) < \tau_{\text{prev}}$ else $I(i,j) \in \Omega^{\text{B}}$.

3. Compute the average of each set as:

$$f = \frac{\sum_{i,j \in \Omega^{\text{F}}} I(i,j)}{|\Omega^{\text{F}}|}, \text{and}$$

$$b = \frac{\sum_{i,j \in \Omega^{\text{B}}} I(i,j)}{|\Omega^{\text{B}}|},$$

   where the image patch is composed of the union of the current foreground and background. Here $|\Omega^{\cdot}|$ is the number of elements in the set $\Omega^{\cdot}$.

4. Update the threshold value

$$\tau_{\text{next}} = \frac{f+b}{2}.$$

5. If $|\tau_{\text{next}} - \tau_{\text{prev}}| > \epsilon$: $\tau_{\text{prev}} = \tau_{\text{next}}$ and go to 2, else return $\tau = \tau_{\text{next}}$.

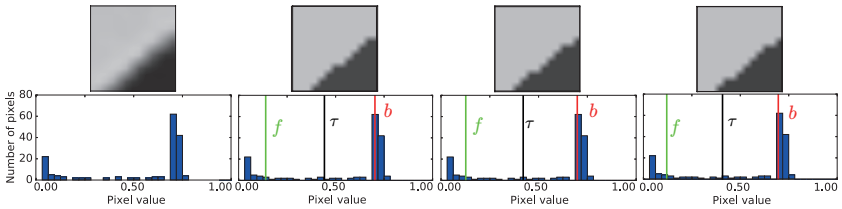

Figure D.1: Example of the iterative process performed by `ISODATA` algorithm. From left to right: input patch, first, second and third (last) iteration of `ISODATA` algorithm, respectively.

The output of the `ISODATA` algorithm is the threshold τ indicating the mid-point between foreground and background modes of the histogram. Figure D.1 shows an example of the `ISODATA` algorithm iteration process for an edge patch. Note that in each iteration the image patch has been labeled by using the average value of the F and B. We use this algorithm for computing the threshold because of its simplicity and accuracy in computing the mid-value between the modes of a histogram with bimodal distribution. Note that the pixel values are assumed in the range of 0 to 1.

Bibliography

(1995). Recommendation ITU-R BT.601-5 studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.601-5-199510-S!!PDF-E.pdf.

(1996). ITU-T recommendation P.930 principles of a reference impairment system for video. ITU, Geneva, Switzerland.

(1998). ITU-R recommendation BT.500-11 methodology for the subjective assessment of the quality of television pictures. ITU, Geneva, Switzerland.

(2000). FRTV video quality database phase I. http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx.

(2004). CIE-159:2004 technical report: A colour appearance model for color management systems CIECAM02. http://div8.cie.co.at/?i_ca_id=623&pubid=200.

Abbadeni, N. (2010). Texture representation and retrieval using the causal autoregressive model. *Visual Communication and Image Representation*, 21:651 – 664.

Abril, H., Millan, M., Torres, Y., and Navarro, R. (1998). Automatic method based on image analysis for pilling evaluation in fabrics. *Journal of Optical Engineering*, 37:2937 1 – 11.

Ade, F. (1983). Characterization of textures by eigenfilters. *Signal Processing*, 5:451 – 457.

Agaian, S., Lentz, K., and Grigoryan, A. (2000). A new measure of image enhancement. In *Proc. of the International Conference on Signal Processing and Communication*, pages 19 – 22.

Agaian, S., Silver, B., and Panetta, K. (2007). Transform coefficient histogram-based image enhancement algorithms using contrast entropy. *IEEE Transactions on Image Processing*, 16:741 – 758.

Ahar, A., Barri, A., and Schelkens, P. (2018). From sparse coding significance to perceptual quality: A new approach for image quality assessment. *IEEE Transactions on Image Processing*, 27:879 – 893.

Aibara, T., Mabuchi, T., and Ohue, K. (1999). Automatic evaluation of the appearance of seam puckers on suits. In *Proc. SPIE 3652, Machine Vision Applications in Industrial Inspection VII*, pages 1 – 4.

Aptoula, E. and Lefevre, S. (2011). *Advances in Imaging and Electron Physics*, chapter Morphological texture description of grayscale and color images, pages 1 – 74. Elsevier.

Areepongsa, S., Park, D., and Rao, K. (2000). Invariant features for texture image retrieval using steerable pyramid. In *Proc. of the 5th Internationa Symposium on Wireless Personal Multimedia Communications*, pages 1 – 5.

Aroussi, M., Hassouni, M., Ghouzali, S., Rziza, M., and Aboutajdine, D. (2009). Novel face recognition approach based on steerable pyramid feature extraction. In *Proc. of the 16th IEEE International Conference on Image Processing*, pages 4165 – 4168.

Babacan, S., Molina, R., and Katsaggelos, A. (2008). Generalized gaussian markov random field image restoration using variational distribution approximation. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1265 – 1268.

Baik, S. and Pachowicz, P. (2002). Application of adaptive object recognition approach to aerial surveillance. In *Proc. of the Fifth International Conference on Information Fusion*, pages 658 – 663.

Bando, E., Hardeberg, J., and Connah, D. (2005). Can gamut mapping quality be predicted by colour image difference formulae? In *Proc. of the IS&T/SPIE Electronic Imaging: Human Vision and Electronic Imaging X*, pages 180 – 191.

Baranczuk, Z., Zolliker, P., and Giesen, J. (2010). Image-individualized gamut mapping algorithms. *Journal of Imaging Science and Technology*, 54:030201 1 – 7.

Barron, J., Fleet, D., Beauchemin, S., and Burkitt, T. (1992). Performance of optical flow techniques. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 236 – 242.

Benjelil, M., Kanoun, S., Mullot, R., and Alimi, A. (2009). Steerable pyramid based complex documents images segmentation. In *Proc. of the 10th International Conference on Document Analysis and Recognition*, pages 833 – 837.

Bertalmio, M., Caselles, V., and Provenzi, E. (2009). Issues about retinex theory and contrast enhancement. *International Journal of Computer Vision*, 83:101 – 119.

Bloomfield, P. and Steiger, W. (1980). Least absolute deviations curve-fitting. *SIAM Journal on Scientific Computing*, 1:290 – 301.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2009). *Introduction to meta - analysis*, chapter Effect sizes based on correltions, pages 41 – 43. John Wiley & Sons, Inc.

Brun, L. and Tremeau, A. (2002). *Digital Color Imaging Handbook*, chapter Color quantization. CRC press.

Burt, P. and Adelson, E. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31:532 – 540.

Busch, A. and Boles, W. (2002). Texture classification using multiple wavelet analysis. In *Proc. of the Digital Image Computing Techniques and Applications*, pages 1 – 5.

Chan, W., Law, N., and Siu, W. (2003). Multiscale feature analysis using directional filter bank. In *Proc. of the Fourth International Joint Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia*, pages 822 – 826.

Chen, C. and Huang, C. (1993). Markov random fields for texture classification. *Pattern Recognition Letters*, 14:907 – 914.

Chen, P. and Popovich, P. (2002a). *Correlation: Parametric and nonparametric measures*, chapter The Pearson Product-Moment Correlation. Sage Publications.

Chen, P. and Popovich, P. (2002b). *Correlation: Parametric and nonparametric measures*, chapter Special Cases of Pearson's R. Sage Publications.

Chen, P. and Popovich, P. (2002c). *Correlation: Parametric and nonparametric measures*, chapter Other Useful Nonparametric Correlations. Sage Publications.

Chen, Y. (1995). *Novel Techniques for image texture classification.* PhD thesis, University of Southampton.

Chikkerur, S., Sundaram, V., Reisslein, M., and Karam, L. (2011). Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Transactions on broadcasting*, 57:165 – 182.

Choi, C., Kim, H., Jin, Y., and Kim, H. (2009). Objective wrinkle evaluation system of fabrics based on 2d fft. *Fibers and Polymers*, 10:260 – 265.

Chou, C. and Liu, K. (2007). A fidelity metric for assessing visual quality of color images. In *Proc. of the International Conference on Computer Communications and Networks*, pages 1154 – 1159.

Cristobal, G. and Hormigo, J. (1999). Texture segmentation through eigen-analysis of the pseudo-wigner distribution. *Pattern Recognition Letters*, 20:337 – 345.

Cross, G. and Jain, A. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:25 – 39.

Daly, S. (1992). Visible differences predictor: an algorithm for the assessment of image fidelity. In *Proc. SPIE 1666, Human Vision, Visual Processing, and Digital Display III*, pages 2 – 16.

Davies, S. and Hall, P. (1999). Fractal analysis of surface roughness by using spatial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):3 – 37.

De-Simone, F., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S., and Ebrahimi, T. (2009). EPFL-PoliMI video quality assessment database. `http://vqa.como.polimi.it/`.

Deguchi, K. (1986). Two-dimensional auto-regressive model for analysis and synthesis of gray-level textures. In *Proc. of the First International Symposium for Science on Form*, pages 441 – 449.

Deng, Y., Kenney, C., Moore, M., and Manjunath, B. (1999). Peer group filtering and perceptual color image quantization. In *Proc. of the IEEE International Symposium on Circuits and Systems*, pages 21 – 24.

Dianat, R. and Kasaei, S. (2008). On automatic threshold selection in regression method for change detection in remote sensing images. In *Proc. of the International Symposium on Telecommunications*, pages 1 – 6.

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *Stata Journal*, 15:292 – 300.

Dippel, S., Stahl, M., Wiemker, R., and Blaffert, T. (2002). Multiscale contrast enhancement for radiographies: Laplacian pyramid versus fast wavelet transform. *IEEE Transactions on Medical Imaging*, 21:343 – 353.

Do, M. and Vetterli, M. (2003). The finite ridgelet transform for image representation. *IEEE Transactions on Image Processing*, 12:16 – 28.

Do, M. and Vetterli, M. (2005). The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14:2091 – 2106.

Dong, Y. and Ma, J. (2011). Wavelet-based image texture classification using local energy histograms. *IEEE Signal Processing Letters*, 18:247 – 250.

Donoho, D. (1999). Wedgelets: Nearly minimax estimation of edges. *The Annals of Statistics*, 27:859 – 897.

Dragoman, D. (2005). Applications of the wigner distribution function in signal processing. *EURASIP Applied Signal Processing*, 2005:1520 – 1534.

Eckstein, W. (1996). Segmentation and texture analysis. In *Proc. of the International Archives of Photogrammetry and Remote Sensing*, pages 165 – 175.

Ejembi, O. and Bhatti, S. (2015). Client-side energy costs of video streaming. In *Proc. of the IEEE International Conference on Data Science and Data Intensive Systems*, pages 252 – 259.

Elunai, R., Chandran, V., and Mabukwa, P. (2010). Digital image processing techniques for pavement macro-texture analysis. In *Proc. of the 24th ARRB Conference: Building on 50 years of road transport research*, pages 1 – 5.

Fairchild, M. (2013). *Color Appearance Models*, chapter Color Appearance Models. John Wiley & Sons, Inc.

Fairchild, M. and Johnson, G. (2004). iCAM framework for image appearance, differences, and quality. *Journal of Electronic Imaging*, 13:126 – 138.

Fan, G. and Xia, X. (2003). Wavelet-based texture analysis and synthesis using hidden markov models. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 50:106 – 120.

Feghali, R., Speranza, F., Wang, D., and Vincent, A. (2007). Video quality metric for bit rate control via joint adjustment of quantization and frame rate. *IEEE Transactions on Broadcasting*, 53:441 – 446.

Fehr, J. and Burkhardt, H. (2008). 3d rotation invariant local binary patterns. In *Proc. of the International Conference on Pattern Recognition*, pages 1 – 4.

Fezza, S., Larabi, M., and Faraoun, K. (2014). Feature-based color correction of multiview video for coding and rendering enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 24:1486 – 1498.

Freeman, W. and Adelson, E. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:891 – 906.

Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*, chapter Similarity and Dissimilarity Measures, pages 67 – 106. ASA-SIAM.

Gao, C., Panetta, K., and Agaian, S. (2013). No reference color image quality measures. In *Proc. of the IEEE International Conference on Cybernetics*, pages 243 – 248.

Gao, H., Miao, Q., Yang, J., and Ma, Z. (2018). Image quality assessment using image description in information theory. *IEEE Access*, Early Access:1 – 13.

Garcia, M., Schleicher, R., and Raake, A. (2010a). Towards a content-based parametric video quality model for IPTV. In *Proc. of the International Workshop on Perceptual Quality of Systems*, pages 1 – 5.

Garcia, S., Fernandez, A., Luengo, J., and Herrera, F. (2010b). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Journal of Electronic Imaging*, 19:011006 1 – 21.

Garcia-Freitas, P., Akamine, W., and Farias, M. (2018). Using multiple spatio-temporal features to estimate video quality. *Signal Processing: Image Communication*, 64:1 – 10.

Gislason, A., Davies, A., and Cowen, A. (2010). Dose optimization in pediatric cardiac x-ray imaging. *Medical Physics*, 37:5258 – 5269.

Goossens, B. (2006). Advanced noise reduction in the wavelet domain for colour video. Master's thesis, Ghent University.

Goossens, B., Aelterman, J., Luong, H., Pizurica, A., and Philips, W. (2009). Efficient design of a low redundant discrete shearlet transform. In *Proc. of the International Workshop on Local and Non-Local Approximation in Image Processing*, pages 112 – 124.

Goossens, B., De-Vylder, J., and Philips, W. (2014). Quasar - a new heterogeneous programming framework for image and video processing algorithms on cpu and gpu. In *Proc. of the IEEE International Conference on Image Processing*, pages 2183 – 2185.

Grath, M. (2003). Markov random field image modelling. Master's thesis, University of Cape Town.

Gu, K., Li, L., Lu, H., Min, X., and Lin, W. (2017). A fast reliable image quality predictor by fusing micro- and macro-structures. *IEEE Transactions on Industrial Electronics*, 64:3903 – 3912.

Gu, K., Qiao, J., Min, X., Yue, G., Lin, W., and Thalmann, D. (2018). Evaluating quality of screen content images via structural variation analysis. *IEEE Transactions on Visualization and Computer Graphics*, Early Access:1 – 13.

Gu, K., Zhai, G., Yang, X., and Zhang, W. (2014). Deep learning network for blind image quality assessment. In *Proc. of the IEEE International Conference on Image Processing*, pages 511 – 515.

Guha, T., Nezhadarya, E., and Ward, R. (2014). Sparse representation-based image quality assessment. *Signal Processing: Image Communication*, 29:1138 – 1148.

Haakma, R., Jarnikov, D., and van-der Stok, P. (2005). *Dynamic and Robust Streaming in and between Connected Consumer-Electronic Devices*, chapter Perceived Quality of Wirelessly Transported Videos, pages 213 – 239. CRC press.

Habekost, M. (2013). Which color differencing equation should be used? *International Circular of Graphic Education and Research*, 6:20 – 33.

Han, Y. and Shi, P. (2007). An adaptive level-selecting wavelet transform for texture defect detection. *Image and Vision Computing*, 25:1239 – 1248.

Haralick, R. (1979). Statistical and structural approaches to texture. *Proc. of the IEEE*, 67:786 – 804.

Hardeberg, J., Bando, E., and Pedersen, M. (2008). Evaluating colour image difference metrics for gamut-mapped images. *Coloration Technology*, 124:243 – 253.

Hasler, D. and Susstrunk, S. (2003). Measuring colourfulness in natural images. In *Proc. of the IS&T/SPIE Electronic Imaging: Human Vision and Electronic Imaging VIII*, pages 87 – 95.

Hassan, M. and Bhagvati, C. (2012). Structural similarity measure for color images. *International Journal of Computer Applications*, 43:7 – 12.

He, L., Gao, X., Lu, W., Li, X., and Tao, D. (2011). Image quality assessment based on S-CIELAB model. *Signal, Image and Video Processing*, 5:283 – 290.

Heeger, D. and Bergen, J. (1995). Pyramid-based texture analysis/synthesis. In *Proc. of the IEEE International Conference on Image Processing*, pages 229 – 238.

Heilbronner, R. (1992). The autocorrelation function: an image processing tool for fabric analysis. *Tectonophysics*, 212:351 – 370.

Hiremath, P. and Pujari, J. (2007). Content based image retrieval using color, texture and shape features. *International journal of computer science and security*, 1:22 – 35.

Hobfeld, T., Schatz, R., and Egger, S. (2011). Sos: The mos is not enough! In *Proc. of the International Workshop on Quality of Multimedia Experience*, pages 131 – 136.

Homigo, J. and Cristobal, G. (1998). High resolution spectral analysis of images using the pseudo-wigner distribution. *IEEE Transactions on Signal Processing*, 46:1757 – 1763.

Hong, G. and Luo, M. (2006). A new algorithm for calculating perceived colour difference of images. *Imaging Science Journal*, 54:1 – 15.

Hu, J. (2008). *Fabric Testing*, chapter Introduction to fabric testing, pages 1 – 26. Woodhead Publishing.

Huertas, R., Melgosa, M., and Oleari, C. (2006). Performance of a color-difference formula based on OSA-UCS space using small-medium color differences. *Journal of the Optical Society of America A*, 23:2077 – 2084.

Huynh-Thu, Q. and Ghanbari, M. (2008). Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800 – 801.

Imai, F., Tsumura, N., and Miyake, Y. (2001). Perceptual color difference metric for complex images based on mahalanobis distance. *Journal of Electronic Imaging*, 10:385 – 393.

Jain, A. and Farrokhnia, F. (1990). Unsupervised texture segmentation using gabor filters. In *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, pages 14 – 19.

Jensen, K. and Carstensen, M. (2002). Fuzz and pills evaluated on knitted textiles by image analysis. *Textile Research Journal*, 72:34 – 50.

Johnson, G. (2006). Using color appearance in image quality metrics. In *Proc. of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pages 1 – 4.

Johnson, G. and Fairchild, M. (2002). *Digital Color Imaging Handbook*, chapter Visual psychophysics and color appearance. CRC press.

Johnson, G. and Fairchild, M. (2003). A top down description of S-CIELAB and CIEDE2000. *Color Research and Application*, 28:425 – 435.

Johnson, G., Song, X., Montag, E., and Fairchild, M. (2010). Derivation of a color space for image color difference measurement. *Color Research and Application*, 35:387 – 400.

Jose, D., Hollies, N., and Spivak, S. (1986). Instrumental techniques to quantify textural change in carpet part i: Image analysis. *Textile Research Journal*, 56:591 – 597.

Joshi, M., Bartakke, P., and Sutaone, M. (2009). Texture representation using autoregressive models. pages 386 – 390.

Kang, T., Kim, S., Sul, I., Youn, J., and Chung, K. (2005). Fabric surface roughness evaluation using wavelet-fractal method part i: Wrinkle, smoothness and seam pucker. *Textile Research Journal*, 75:751 – 760.

Kashyap, R. and Chellappa, R. (1983). Estimation and choice of neighbors in spatial-interaction models of images. *IEEE Transactions on Information Theory*, 29:60 – 72.

Kaur, M., Kumar, D., Walia, E., and Sandhu, M. (2011). Periodic noise removal in strain and natural images using 2-d fast fourier transform. *International Journal of Computer and Communication Technology*, 2:1 – 5.

Keimel, C., Oelbaum, T., and Diepold, K. (2009). Improving the verification process of video quality metrics. In *Proc. of the International Workshop on Quality of Multimedia Experience*, pages 121 – 126.

Keimel, C., Oelbaum, T., and Diepold, K. (2010). Improving the prediction accuracy of video quality metrics. In *Proc. of the International Conference on Acoustics Speech and Signal Processing*, pages 2442 – 2445.

Khan, A., Lingfen, S., and Ifeachor, E. (2009). Content clustering based video quality prediction model for mpeg4 video streaming over wireless networks. In *Proc. of the International Conference on Communications*, pages 1 – 5.

Kim, J. and Lee, S. (2017). Deep learning of human visual sensitivity in image quality assessment framework. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1676 – 1684.

Kingdom, F. and Prins, N. (2010a). *Psychophysics: A practical introduction*, chapter Classifying Psychophysical Experiments, pages 9 – 37. Pearson Prentice Hall.

Kingdom, F. and Prins, N. (2010b). *Psychophysics: A practical introduction*, chapter Psychometric Functions, pages 59 – 120. Pearson Prentice Hall.

Kivinen, H., Nuutinen, M., and Oittinen, P. (2010). Comparison of colour difference methods for natural images. In *Proc. of the Conference on Colour in Graphics, Imaging, and Vision*, pages 510 – 515.

Korhonen, J. and You, J. (2010). Improving objective video quality assessment with content analysis. In *Proc. of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pages 1 – 6.

Krasula, L., Klima, M., Rogard, E., and Jeanblanc, E. (2011). Matlab-based applications for image processing and image quality assessment - part i: Software description. In *Proc. of Czech and Slovak Technical Universities*, pages 1009 – 1015.

Krishnamachari, S. and Chellappa, R. (1997). Multiresolution gauss-markov random field models for texture segmentation. *IEEE Transactions on Image Processing*, 6:251 – 267.

Krishnamoorthi, R. and Seetharaman, K. (2007). Image compression based on a family of stochastic models. *Signal Processing*, 87:408 – 416.

Kumcu, A., Ortiz-Jaramillo, B., Platisa, L., Goossens, B., and Philips, W. (2015a). Interventional x-ray quality measure based on a psychovisual detectability model. In *Proc. of the Medical Image Perception Conference XVI*, pages 37 – 37.

Kumcu, A., Platisa, L., Chen, H., Gislason-Lee, A., Davies, A., Schelkens, P., Taeymans, Y., and Philips, W. (2015b). Selecting stimuli parameters for video quality assessment studies based on perceptual similarity distances. In *Proc. SPIE 9399, Image Processing: Algorithms and Systems XIII*, pages 93990F 1 – 10.

Lam, E. (2008). Texture classification using wavelet decomposition. In *Proc. of the IEEE International Conference on System of Systems Engineering*, pages 1 – 5.

Larson, E. and Chandler, D. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19:011006 1 – 21.

Lavvafi, M., Monadjemi, S., and Moallem, P. (2010). Film colorization, using artificial neural networks and laws filters. *Journal of Computers*, 5:1094 – 1099.

Laws, K. (1980). *Textured Image Segmentation*. PhD thesis, University of Southern California Los Angeles Image Processing Inst.

Le-Callet, P., Pechard, S., Tourancheau, S., Ninassi, A., and Barba, D. (2007). Towards the next generation of video and image quality metrics: impact of display, resolution, contents and visual attention in subjective assessment. In *Proc. of the International Workshop on Image Media Quality and its Applications*, pages 1 – 9.

Lee, C., Seo, G., and Lee, S. (2009). Epsnr for objective image quality measurements. In *Proc. of the International Conference on Computer Imaging Theory and Applications*, pages 92 – 95.

Lee, D. and Rogers, E. (2014). Towards a novel perceptual color difference metric using circular processing of hue components. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 166 – 170.

Lee, K. and Chen, L. (2002). A new method for coarse classification of textures and class weight estimation for texture retrieval. *Pattern Recognition and Image Analysis*, 12(4):400 – 410.

Lee, S., Xin, J., and Westland, S. (2005). Evaluation of image similarity by histogram intersection. *Color Research and Application*, 30:265 – 274.

Li, J., Koudota, Y., Barkowsky, M., Primon, H., and Le-Callet, P. (2014). Comparing upscaling algorithms from hd to ultra hd by evaluating preference of experience. In *Proc. of the International Workshop on Quality of Multimedia Experience*, pages 208 – 213.

Li, L., Xia, W., Lin, W., Fang, Y., and Wang, S. (2017). No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features. *IEEE Transactions on Multimedia*, 19:1030 – 1040.

Li, S., Ma, L., and Ngan, K. (2011a). Image quality assessment by separately evaluating detail losses and additive impairments. `http://www.ee.cuhk.edu.hk/~snli/VQM.htm`.

Li, S., Ma, L., and Ngan, K. (2011b). Video quality assessment by decoupling additive impairments and detail losses. In *Proc. of the International Workshop on Quality of Multimedia Experience*, pages 90 – 95.

Lindeberg, T. and Eklundh, J. (1992). Scale-space primal sketch: construction and experiments. *Image and Vision Computing*, 10(1):3 – 18.

Lissner, I., Preiss, J., Urban, P., Lichtenauer, M.-S., and Zolliker, P. (2013). Image-difference prediction: From grayscale to color. *IEEE Transactions on Image Processing*, 22:435 – 446.

Liu, H., Huang, M., Cui, G., Luo, M., and Melgosa, M. (2013a). Color-difference evaluation for digital images using a categorical judgment method. *Journal Optical Society of America A*, 30:616 – 626.

Liu, H., Huang, M., Xie, M., and Zhu, M. (2010). The effect of image's contents on the visual color difference for digital images. In *Proc. of the International Congress on Image and Signal Processing*, pages 2356 – 2359.

Liu, T., Lin, Y., Lin, W., and Kuo, C. (2013b). Visual quality assessment: recent developments, coding applications and future trends. *APSIPA Transactions on Signal and Information Processing*, 2:1 – 20.

Liu, X. and Wang, D. (2003). Texture classification using spectral histograms. *IEEE Transactions On Image Processing*, 12:661 – 670.

Lukin, V., Ponomarenko, N., Ieremeiev, O., Egiazarian, K., and Astola, J. (2015). Combining full-reference image visual quality metrics by neural network. In *Proc. SPIE 9394, Human Vision and Electronic Imaging XX*, pages 93940K 1 – 12.

Luo, M., Cui, G., and Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research and Application*, 26:340 – 350.

Lv, Y., Jiang, G., Yu, M., Xu, H., Shao, F., and Liu, S. (2015). Difference of gaussian statistical features based blind image quality assessment: A deep learning approach. In *Proc. of the IEEE International Conference on Image Processing*, pages 2344 – 2348.

Ly, D., Beucher, S., Bilodeau, M., Persa, S., Damstra, K., Pot, R., and Rooy, J. (2015). Automatic color correction: region-based approach and performance evaluation using full reference metrics. *Journal of Electronic Imaging*, 24:061207 1 – 9.

Ly, D., S.Beucher, and M.Bilodeau (2014). Color correction through region matching leveraged by point correspondences. In *Proc. of the IEEE International Conference on Image Processing*, pages 640 – 644.

Ma, W. and Manjunath, B. (1995). A comparison of wavelet transform features for texture image annotation. In *Proc. of the International Conference on Image Processing*, pages 256 – 259.

Maenpaa, T. (2003). *The Local Binary Pattern Approach To Texture Analysis: Extensions And Applications*. PhD thesis, University of Oulu.

Mak, K. and Li, W. (2008). Objective evaluation of seam pucker on textiles by using self-organizing map. *International Journal of Computer Science*, 35:1 – 8.

Mallat, S. (2009). *A Wavelet Tour of Signal Processing: The Sparse Way*, chapter Sparse Representations, pages 1 – 31. Academic Press.

Mandic, L., Grgic, S., and Grgic, M. (2006). Comparison of color difference equations. In *Proc. of the International Symposium focused on Multimedia Signal Processing and Communications*, pages 107 – 110.

Manjunath, B. and Ma, W. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:837 – 842.

Mao, J. and Jain, A. (1992). Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition Society*, 25:173 – 188.

Massart, D., Smeyers-Verbeke, J., Caprona, X., and Schlesierb, K. (2005). Visual presentation of data by means of box plots. *LC GC Europe*, 18:2 – 5.

Mazher, A. and Ali, A. (2011). Texture analysis of brodatz images using statistical methods. *Engineering and Technology*, 29:716 – 724.

Mendes, A., Fiadeiro, P., and Lopes, R. (2010). Subjective and objective pilling evaluations of textile fabrics: a comparison. *Textile Research Journal*, 80:1887 – 1897.

Menor, D., Mello, C., and Zanchettin, C. (2016). Objective video quality assessment based on neural networks. *Procedia Computer Science*, 96:1551 – 1559.

Militký, J. and Bleša, M. (2008). Evaluation of patterned fabric surface roughness. *Indian Journal of Fibre and Textile Research*, 33:246 – 252.

Ming, Y., Huijuan, L., Yingchun, G., and Dongming, Z. (2009). A method for reduced-reference color image quality assessment. In *Proc. of the International Congress on Image and Signal Processing*, pages 1 – 5.

Mohri, M., Hosseini-Ravandim, S., and Youssefi, M. (2005). Objective evaluation of wrinkled fabric using radon transform. *Journal of the Textile Institute*, 96:365 – 370.

Mojsilovic, A., Jianying, H., and Soljanin, E. (2002). Extraction of perceptually important colors and similarity measurement for image matching, retrieval and analysis. *IEEE Transactions on Image Processing*, 11:1238 – 1248.

Mojsilovic, A., Popovic, M., and Rackov, D. (2000). On the selection of an optimal wavelet basis for texture characterization. *IEEE Transactions on Image Processing*, 9:2043 – 2050.

Monadjemi, A., Mirmehdi, M., and Thomas, B. (2004). Restructured eigenfilter matching for novelty detection in random textures. In *Proc. of the 15th British Machine Vision Conference*, pages 637 – 646.

Moorthy, A. and Bovik, A. (2009a). Perceptually significant spatial pooling techniques for image quality assessment. In *Proc. SPIE 7240, Human Vision and Electronic Imaging XIV*, pages 1 – 11.

Moorthy, A. and Bovik, A. (2009b). Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3:193 – 201.

Moorthy, A. and Bovik, A. (2010). Efficient video quality assessment along temporal trajectories. *IEEE transactions on circuits and systems for video technology*, 20:1653 – 1658.

Mori, T. and Komiyama, J. (2002). Evaluating wrinkled fabrics with image analysis and neural networks. *Textile Research Journal*, 72:417 – 422.

Moroney, N., Fairchild, M., Hunt, R., Li, C., Luo, M., and Newman, T. (2002). The CIECAM02 color appearance model. In *Proc. of the IS&T/SID Color Imaging Conference*, pages 23 – 27.

Morovic, J. (2008). *Color Gamut Mapping*, chapter Desired Color Reproduction Properties and their Evaluation. John Wiley & Sons, Inc.

Muijs, R. and Kirenko, I. (2005). A no-reference blocking artifact measure for adaptive video processing. In *Proc. of the European Signal Processing Conference*, pages 1 – 4.

Mukaka, M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24:69 – 71.

Murthy, A. and Karam, L. (2010). A matlab-based framework for image and video quality evaluation. In *Proc. of the Conference on Quality of Multimedia Experience*, pages 242 – 247.

Na, Y. and Pourdeyhimi, B. (1995). Assessing wrinkling using image analysis and replicate standards. *Textile Research Journal*, 65:149 – 157.

Nixon, M. and Aguado, A., editors (2002). *Feature Extraction and Image Processing*, chapter Introduction to texture description, segmentation and classification, pages 291 – 310. Newnes.

Ojala, T., Pietikainen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on feature distribution. *Pattern Recognition*, 29:51 – 59.

Oleari, C. (2004). Color opponencies in the system of the uniform color scales of the optical society of america. *Journal of the Optical Society of America A*, 21:677 – 682.

Oleari, C., Melgosa, M., and Huertas, R. (2008). Euclidean color-difference formula for small-medium color differences in log-compressed OSA-UCS space. *Journal of the Optical Society of America A*, 26:121 – 134.

Orjuela-Vargas, S. (2012). *Texture analysis for the evaluation of appearance changes in textile surfaces*. PhD thesis, Ghent University.

Orjuela-Vargas, S., Vansteenkiste, E., Rooms, F., De-Meulemeester, S., De-Keyser, R., and Philips, W. (2010). Evaluation of the wear label description in carpets by using local binary pattern techniques. *Textile Research Journal*, 80(20):2132 – 2143.

Orjuela-Vargas, S., Vansteenkiste, E., Rooms, F., Philips, W., De-Keyser, R., and Van-Langenhove, L. (2008). Carpet wear classification using coocurrence matrices and support vector machines. In *Proc. of the Annual Workshop on Circuits, Systems and Signal Processing*, pages 378 – 383.

Ortiz-Jaramillo, B. (2017a). Control-aware video quality estimation - smart media. `http://www.itf2017.be/Program/Experts/page.aspx/2546`.

Ortiz-Jaramillo, B. (2017b). iFAS: image fidelity assessment. `https://telin.ugent.be/~bortiz/ifas`.

Ortiz-Jaramillo, B., Garcia-Alvarez, J., Orjuela-Vargas, S., Fuhr, H., Castellanos-Dominguez, G., and Philips, W. (2012). Quantifying image distortion based on gabor filter bank and multiple regression analysis. In *Proc. SPIE 8293, Image Quality and System Performance IX*, pages 82930E 1 – 10.

Ortiz-Jaramillo, B., Kumcu, A., and Philips, W. (2016a). Evaluating color difference measures in images. In *Proc. of the Conference on Quality of Multimedia Experience*, pages 1 – 6.

Ortiz-Jaramillo, B., Kumcu, A., Platisa, L., and Philips, W. (2014a). A full reference video quality measure based on motion differences and saliency maps evaluation. In *Proc. of the International Conference on Computer Vision Theory and Applications*, pages 714 – 722.

Ortiz-Jaramillo, B., Kumcu, A., Platisa, L., and Philips, W. (2015a). Computing contrast ratio in images using local content information. In *Proc. of the Symposium on Signal Processing, Images and Computer Vision*, pages 1 – 6.

Ortiz-Jaramillo, B., Kumcu, A., Platisa, L., and Philips, W. (2015b). Computing contrast ratio in medical images using local content information. In *Proc. of the Medical Image Perception Conference XVI*, pages 33 – 33.

Ortiz-Jaramillo, B., Kumcu, A., Platisa, L., and Philips, W. (2018a). Content-aware contrast ratio measure for images. *Journal of Signal Processing: Image Communication*, 62:51 – 63.

Ortiz-Jaramillo, B., Kumcu, A., Platisa, L., and Philips, W. (2018b). Evaluation of color differences in natural scene color images. *Journal of Signal Processing: Image Communication*, Submitted.

Ortiz-Jaramillo, B., Nino-Castaneda, J., Platisa, L., and Philips, W. (2015c). Content-aware video quality assessment: predicting human perception of quality using peak signal to noise ratio and spatial/temporal activity. In *Proc. SPIE 9399, Image Processing: Algorithms and Systems XIII*, pages 939917 1 – 12.

Ortiz-Jaramillo, B., Nino-Castaneda, J., Platisa, L., and Philips, W. (2016b). Content-aware objective video quality assessment. *Journal of Electronic Imaging*, 25:013011 1 – 16.

Ortiz-Jaramillo, B., Orjuela-Vargas, S., Van-Langenhove, L., Castellanos-Dominguez, C., and Philips, W. (2014b). Reviewing, selecting and evaluating features in distinguishing fine changes of global texture. *Pattern Analysis And Applications*, 17:1 – 15.

Ortiz-Jaramillo, B., Platisa, L., and Philips, W. (2017). ifas : image fidelity assessment. In *Proc. of the International Workshop Computational Color Imaging*, pages 83 – 94.

Ou, Y., Ma, Z., Liu, T., and Wang, Y. (2011). Perceptual quality assessment of video considering both frame rate and quantization artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, 21:286 – 298.

Ou, Y., Xue, Y., and Wang, Y. (2014). Q-star: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions. *IEEE Transactions on Image Processing*, 23:2473 – 2486.

Ouni, S., Zagrouba, E., Chambah, M., and Herbin, M. (2008). A new spatial colour metric for perceptual comparison. In *Proc. of the International Conference on Computing and e-Systems*, pages 413 – 428.

Paget, R. and Longstaff, D. (1995). Texture synthesis via a non-parametric markov random field. In *Proc. of the Digital Image Computing: Techniques and Applications*, pages 547 – 552.

Palma-Amestoy, R., Provenzi, E., Bertalmio, M., and Caselles, V. (2009). A perceptually inspired variational framework for color enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:458 – 474.

Palmer, S., Zhang, J., and Wang, X. (2009). New methods for objective evaluation of fabric pilling by frequency domain image processing. *Research journal of textile and apparel*, 13:11 – 23.

Panetta, K., Gao, C., and Agaian, S. (2013). No reference color image contrast and quality measures. *IEEE Transactions on Consumer Electronics*, 59:643 – 651.

Pappas, T., Safranek, R., and Chen, J. (2010). *Handbook of Image and Video Processing*, chapter Perceptual Criteria for Image Quality Evaluation. Academic Press.

Park, W., Madden, D., Rockmore, D., and Chirikjian, G. (2010). Deblurring of class-averaged images in single-particle electron microscopy. *Inverse Problems*, 26:1 – 29.

Pechard, S., Pepion, R., and Le-Callet, P. (2011). IRCCyN IVC 1080i: an HD video quality database. `www.irccyn.ec-nantes.fr/spip.php?article541`.

Pedersen, M. and Hardeberg, J. (2009). A new spatial hue angle metric for perceptual image difference. In *Proc. of the Computational Color Imaging*, pages 81 – 90.

Pedersen, M., Rizzi, A., Hardeberg, J., and Simone, G. (2008). Evaluation of contrast measures in relation to observers perceived contrast. In *Proc. of the European Conference on Colour in Graphics, Imaging, and Vision*, pages 253 – 258.

Peli, E. (1997). In search of a contrast metric: Matching the perceived contrast of gabor patches at different phases and bandwidths. *Vision Research*, 37:3217 – 3224.

Petrou, M. and Sevilla, P., editors (2006). *Image Processing: Dealing With Textures*, chapter Stationary grey texture images, pages 81 – 295. John Wiley and Sons, Ltd.

Pfister, R., Schwarz, K., Janczyk, M., Dale, R., and Freeman, J. (2013). Good things peak in pairs: a note on the bimodality coefficient. *Frontiers in Psychology*, 4:1 – 4.

Pico, I., Asensi, C., Crespi, G., Quintanilla, L., and Benavente, M. (2001). A comparative study of texture analysis algorithms in textile inspection applications. In *Proc. of the 12th Scandinavian Conference on Image Analysis*, pages 1 – 4.

Pinson, M. and Wolf, S. (2004a). A new standardized method for objectively measuring video quality. *IEEE transactions on broadcasting*, 50:312 – 322.

Pinson, M. and Wolf, S. (2004b). Video quality metric (vqm) software. `http://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx`.

Pitrey, Y., Barkowsky, M., Pepion, R., Le-Callet, P., and Hlavacs, H. (2012). Influence of the source content and encoding configuration on the perceived quality for scalable video coding. In *Proc. of SPIE-IS&T Electronic Imaging*, pages 82911K 1 – 7.

Platisa, L., Pizurica, A., Vansteenkiste, E., and Philips, W. (2011). No-reference blur estimation based on the average cone ratio in the wavelet domain. In *Proc. of the Multimedia on Mobile Devices and Multimedia Content Access: Algorithms and Systems V*, pages 7881 1 – 12.

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., and Jay-Kuo, C. (2015). Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57 – 77.

Popescu, D., Dobrescu, R., and Nicolae, M. (2007). Texture classification and defect detection by statistical features. *Circuits, Systems and Signal Processing*, 1:79 – 84.

Pourdeyhimi, B., Xu, B., and Nayernouri, A. (1994a). Evaluating carpet appearance loss: Pile lay orientation. *Textile Research Journal*, 64:130 – 135.

Pourdeyhimi, B., Xu, B., and Wehrle, L. (1994b). Evaluating carpet appearance loss: Periodicity and tuft placement. *Textile Research Journal*, 64:21 – 32.

Preiss, J., Fernandes, F., and Urban, P. (2014). Color-image quality assessment: From prediction to optimization. *IEEE Transactions on Image Processing*, 23:1366 – 1378.

Provenzi, E. and Caselles, V. (2014). A wavelet perspective on variational perceptually-inspired color enhancement. *International Journal of Computer Vision*, 106:153 – 171.

Python (2016). Python software foundation. `https://www.python.org/`.

Quinones-Lara, R., Ortiz-Jaramillo, B., Orjuela-Vargas, S., De-Meulemeester, S., Van-Langenhove, L., and Philips, W. (2011). Measuring hairiness in carpets by using surface metrology. In *Proc. SPIE 8133, International Society for Optical Engineering*, pages 81330W 1 – 7.

Radun, J., Nuutinen, M., Antons, J., and Arndt, S. (2017). Did you notice it? how can we predict the subjective detection of video quality changes from eye movements? *IEEE Journal of Selected Topics in Signal Processing*, 11:37 – 47.

Rajashekar, U., Wang, Z., and Simoncelli, E. (2009). Quantifying color image distortions based on adaptive spatio-chromatic signal decompositions. In *Proc. of the IEEE International Conference on Image Processing*, pages 2213 – 2216.

Rajashekar, U., Wang, Z., and Simoncelli, E. (2010). Perceptual quality assessment of color images using adaptive signal representation. In *Proc. of the IS&T/SPIE Electronic Imaging: Human Vision and Electronic Imaging XV*, pages 75271L 1 – 9.

Randen, T. and Husoy, J. (1999). Filtering for texture classification: A comparative study. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 21:291 – 310.

Redondo, R., Fischer, S., Sroubek, F., and Cristobal, G. (2008). A 2d wigner distribution-based multisize windows technique for image fusion. *Visual Communication and Image Representation*, 19:12 – 19.

Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *IEEE Computer Graphics and Applications*, 21:34 – 41.

Rizzi, A., Algeri, T., Medeghini, G., and Marini, D. (2004). A proposal for contrast measure in digital images. In *Proc. of the Conference on Color in Graphics, Imaging and Vision*, pages 1 – 6.

Rodriguez, D., Rosa, R., Costa, E., Abrahao, J., and Bressan, G. (2014). Video quality assessment in video streaming services considering user preference for video content. *IEEE Transactions on Consumer Electronics*, 60:436 – 444.

Ruderman, D., Cronin, T., and Chiao, C. (1998). Statistics of cone responses to natural images: implications for visual coding. *Journal of the Optical Society of America A*, 15:2036 – 2045.

Saint-Marc, P., Chen, J., and Medioni, G. (1991). Adaptive smoothing: a general tool for early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:514 – 529.

Seshadrinathan, K. and Bovik, A. (2010). Motion tuned spatio-temporal quality assessment of natural videos. *IEEE transactions on image processing*, 19:335 – 350.

Seshadrinathan, K., Soundararajan, R., Bovik, A., and Cormack, L. (2010). LIVE video quality database. `http://live.ece.utexas.edu/research/quality/live_video.html`.

Sharma, G. (2002). *Digital Color Imaging Handbook*, chapter Color fundamentals for digital imaging. CRC press.

Sheikh, H., Sabir, M., and Bovik, A. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15:3440 – 3451.

Siew, L., Hodgson, R., and Wood, E. (1988). Texture measures for carpet wear assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:92 – 105.

Silverstein, D. and Farrell, J. (1996). Relationship between image fidelity and image quality. In *Proc. of the International Conference on Image Processing*, pages 881 – 884.

Simoncelli, E. and Freeman, W. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. of the 2nd IEEE International Conference on Image Processing*, pages 444 – 447.

Simone, G., Oleari, C., and Farup, I. (2009). An alternative color difference formula for computing image difference. In *Proc. of the Gjøvik Color Imaging Symposium*, pages 8 – 11.

Singh, M. and Singh, S. (2002). Spatial texture analysis: a comparative study. In *Proc. of the 16th International Conference on Pattern Recognition*, pages 676 – 679.

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349 – 1380.

Smith, A. (1978). Color gamut transform pairs. *Computer Graphics*, 12:12 – 19.

Smith, T. and Guild, J. (1931). The C.I.E. colorimetric standards and their use. *Transactions of the Optical Society*, 33:73 – 134.

Song, Y., Li, Q., Feng, D., Zou, J., and Cai, W. (2016). Texture image classification with discriminative neural networks. *Computational Visual Media*, 2:367 – 377.

Stocker, A. and Simoncelli, E. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9:578 – 585.

Sun, J., Yao, M., Xu, B., and Bel, P. (2011). Fabric wrinkle characterization and classification using modified wavelet coefficients and support-vector-machine classifiers. *Textile Research Journal*, 81:902 – 913.

Suzuki, M., Yaginuma, Y., and Kodama, H. (2009). A 2d texture image retrieval technique based on texture energy filters. In *Proc. of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 145 – 151.

Swain, M. and Ballard, D. (1990). Indexing via color histograms. In *Proc. of the Third International Conference on Computer Vision*, pages 390 – 393.

Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769 – 2794.

Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8:460 – 473.

ASTM-D5684-10 (2010). Standard terminology relating to pile floor coverings. ASTM International.

BSIR (2016). What is interventional radiology? `http://www.bsir.org/`.

Cambridge dictionary (2016). Texture. `http://dictionary.cambridge.org/dictionary/english/texture`.

CIE (1976). Colorimetry - part 4: CIE 1976 L*a*b* colour space. `https://www.iso.org/obp/ui/#iso:std:iso:11664:-4:ed-1:v1:en`.

CRI Test Method 103 (2015). Standard test method for the evaluation of texture appearance retention of carpets for minimum carpet standards program. `http://www.carpet-rug.org/technical-bulletins-and-papers.html#CRI_Laboratory_Test_Methods`.

Cython (2016). C extensions for Python. `http://cython.org/`.

GLib (2016). Python Bindings for GLib/GObject/GIO/GTK+. `https://wiki.gnome.org/Projects/PyGObject`.

ISO-10361:2000 (2005). Textile floor coverings - production of changes in appearance by means of vettermann drum and hexapod tumbler testers. `https://www.iso.org/standard/61692.html`.

ISO-2424:2007 (2007). Textile floor coverings - vocabulary. ISO, Geneva, Switzerland.

ISO 4918:2016 (2016). Resilient, textile and laminate floor coverings - castor chair test. `https://www.iso.org/standard/64610.html`.

ISO 9405:2015 (2015). Textile floor coverings - assessment of changes in appearance. `https://www.iso.org/obp/ui/#iso:std:iso:9405:ed-2:v1:en`.

ITU (2012). Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf.

Kodak (2013). Kodak lossless true color image suite. http://r0k.us/graphics/kodak/.

OpenCV (2016). OpenCV-Python. http://opencv.org/.

PIL (2016). Python Imaging Library (PIL). http://www.pythonware.com/products/pil/.

Pycairo (2016). Cairo graphics library. https://cairographics.org/pycairo/.

PyWavelets (2016). PyWavelets - Wavelet Transforms in Python. https://pywavelets.readthedocs.io/en/latest/.

SciPy (2016). SciPy. https://www.scipy.org/.

TID2013 (2013). Tampere image database 2013 TID2013, version 1.0. http://www.ponomarenko.info/tid2013.htm.

US-FDA (2005). Part iii department of health and human services 21 cfr part 1020 electronic products; performance standard for diagnostic x-ray systems and their major components; final rule. Federal Register vol. 70 (111).

USC Viterbi (2016). The USC-SIPI image database. http://sipi.usc.edu/database/.

Thung, K., Paramesran, R., and Lim, C. (2012). Content-based image quality metric using similarity measure of moment vectors. *Pattern Recognition*, 45:2193 – 2204.

Tkacenko, A. and Vaidyanathan, P. (2003). On the eigenfilter design method and its applications: A tutorial. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 50:497 – 517.

Toet, A. and Lucassen, M. (2003). A new universal colour image fidelity metric. *Displays*, 24:197 – 207.

Topi, M., Timo, O., Matti, P., and Maricor, S. (2000). Robust texture classification by subsets of local binary patterns. In *Proc. of the International Conference on Pattern Recognition*, pages 935 – 938.

Torabi, A., Fossen, H., and Alaei, B. (2008). Application of spatial correlation functions in permeability estimation of deformation bands in porous rocks. *Journal of Geophysical Research*, 113:1 – 10.

Torres-Vega, M., Mocanu, D., Famaey, J., Stavrou, S., and Liotta, A. (2017). Deep learning for quality assessment in live video streaming. *IEEE Signal Processing Letterss*, 24:736 – 740.

Tuceryan, M. and Jain, A. (1998). *The Handbook of Pattern Recognition and Computer Vision*, chapter Texture Analysis, pages 1 – 41. World Scientific Publishing Co.

Vaidya, V. and Padole, C. (2008). Night vision enhancement using wigner distribution. In *Proc. of the 3rd International Symposium on Communications, Control and Signal Processing*, pages 1268 – 1272.

Van-Wallendael, G., Coppens, P., Paridaens, T., Van-Kets, N., den Broeck, W. V., and Lambert, P. (2016). Perceptual quality of 4k-resolution video content compared to hd. In *Proc. of the International Workshop on Quality of Multimedia Experience*, pages 1 – 6.

Video-Quality-Experts-Group (2003). The validation of objective models of video quality assessment, phase II. `https://www.itu.int/md/T01-SG09-C-0060`.

Vo, A., Nguyen, T., and Oraintara, S. (2006). Texture image retrieval using complex directional filter bank. In *Proc. of the IEEE International Symposium on Circuits and Systems*, pages 5495 – 5498.

VQEG (2010). Report on the validation of video quality models for high definition video content. `https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx`.

Waegemana, W., Cottyn, J., Wyns, B., Boullart, L., De-Baets, B., Van-Langenhove, L., and Detand, J. (2008). Classifying carpets based on laser scanner data. *Engineering Applications of Artificial Intelligence*, 21:907 – 918.

Wahba, M., Ashour, A., Napoleon, S., Abd-Elnaby, M., and Guo, Y. (2017). Combined empirical mode decomposition and texture features for skin lesion classification using quadratic support vector machine. *Health Information Science and Systems*, 5:1 – 10.

Wang, J. and Wood, E. (1994). A new method for measuring carpet texture change. *Textile Research Journal*, 65:196 – 202.

Wang, S., Deng, C., Lin, W., Huang, G., and Zhao, B. (2016). Nmf-based image quality assessment using extreme learning machine. *IEEE Transactions on Cybernetics*, 47:232 – 243.

Wang, Z. and Bovik, A. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 3:81 – 84.

Wang, Z. and Bovik, A. (2006). *Modern Image Quality Assessment*, chapter Introduction, pages 1 – 16. Morgan & Claypool Publishers.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600 – 612.

Wang, Z. and Li, Q. (2007). Video quality assessment using a statistical model of human visual speed perception. *Journal of optical society of America*, 24:B 61 – 69.

Wang, Z. and Yong, J. (2008). Texture analysis and classification with linear regression model based on wavelet transform. *IEEE Transactions on Image Processing*, 17:1421 – 11430.

Wei, X. and Mulligan, J. (2010). Performance evaluation of color correction approaches for automatic multi-view image and video stitching. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 263 – 270.

Weszka, J., Dyer, C., and Rosenfeld, A. (1976). A comparative study of texture measures for terrain classification. *IEEE Transactions On Systems, Man, And Cybernetics*, 6(4):269 – 285.

Winkler, S. (2005). *Digital Video Quality: Vision Models and Metrics*, chapter Video Quality, pages 35 – 70. John Wiley & Sons, Inc.

Winkler, S. (2009). On the properties of subjective ratings in video quality experiments. In *Proc. of the International Workshop on Quality of Multimedia Experience*, pages 139 – 144.

Winkler, S. (2010). Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6:616 – 625.

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, chapter Credibility: Evaluating What's Been Learned, pages 147 – 187. Morgan Kaufmann.

Wood, E. (1993). Description and measurement of carpet appearance. *Textile Research Journal*, 63:580 – 594.

Wu, A., Schulze, M., and Castleman, K. (1998). Steerable pyramid filters for selective image enhancement applications. In *Proc. of the IEEE International Symposium on Circuits and Systems*, pages 325 – 328.

Wu, Q., Li, H., Meng, F., and Ngan, K. (2018). A perceptually weighted rank correlation indicator for objective image quality assessment. *IEEE Transactions on Image Processing*, 27:2499 – 2513.

Xiaojun, L., Huabing, H., Yushu, L., and Hong, Z. (2009). The evaluation system of fabric pilling based on image processing technique. In *Proc. of the International Conference on Image Analysis and Signal Processing*, pages 44 – 47.

Xie, X. (2008). A review of recent advances in surface defect detection using texture analysis techniques. *Electronic Letters on Computer Vision and Image Analysis*, 7:1 – 22.

Xin, W., Georganas, N., and Petriu, E. (2011). Fabric texture analysis using computer vision techniques. *IEEE Transactions on Instrumentation and Measurement*, 60:44 – 56.

Xingmei, L., Liang, C., and Jin, W. (2010). The application of laplacian pyramid in image super-resolution reconstruction. In *Proc. of the 2nd International Conference on Signal Processing Systems*, pages 157 – 159.

Xu, B. (1997). Quantifying surface roughness of carpets by fractal dimension. *Clothing and Textiles Research Journal*, 15:155 – 161.

Xu, H., Tong, Q., and Gong, R. (2013). Color difference evaluation for digital images. In *Proc. of the Asia Color Association Conference*, pages 28 – 31.

Yang, F. and Jiang, T. (2003). Pixon-based image segmentation with markov random fields. *IEEE Transactions on Image Processing*, 12:1552 – 1559.

Yang, R., Su, J., and Yu, W. (2017). No-reference image quality assessment based on deep learning method. In *Proc. of the IEEE Information Technology and Mechatronics Engineering Conference*, pages 476 – 479.

Yang, Y., Ming, J., and Yu, N. (2012). Color image quality assessment based on CIEDE2000. *Advances in Multimedia*, pages 1 – 6.

Yazdani, F. and Andani, M. (2017). Verification based on palm vein by estimating wavelet coefficient with autoregressive model. In *Proc. of the Conference on Swarm Intelligence and Evolutionary Computation*, pages 118 – 122.

Yeganeh, H. and Wang, Z. (2013). Subject-rated image database of tone-mapped images. `https://ece.uwaterloo.ca/~z70wang/research/tmqi/`.

Yong, W., Bi, Z., and Gui, S. (2010). A method of generating global view texture images in 3d face modeling. In *Proc. of the 2nd International Conference on Information Science and Engineering*, pages 3435 – 3438.

Zar, J. (2009). *Biostatistical analysis*, chapter Measures of Central Tendency, pages 21 – 30. Pearson Prentice Hall.

Zhang, E., Chen, Y., Yu, J., and Yang, Y. (2010a). A study of image color quality evaluation based on S-CIELAB. In *Proc. of the International Congress on Image and Signal Processing*, pages 1110 – 1114.

Zhang, F., Li, S., Ma, L., Wong, Y., and Ngan, K. (2011). IVP subjective quality video database. `http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtml`.

Zhang, J. and Tan, T. (2002). Brief review of invariant texture analysis methods. *Pattern Recognition*, 35:735 – 747.

Zhang, L., Shen, Y., and Li, H. (2014). Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23:4270 – 4281.

Zhang, L., Zou, B., Zhang, J., and Zhang, Y. (2010b). Classification of polarimetric sar image based on support vector machine using multiple-component scattering model and texture features. *EURASIP Journal on Advances in Signal Processing*, 2010:960831 1 – 9.

Zhang, W. and Liu, H. (2017). Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, 26:1275 – 1288.

Zhang, X. and Wandell, B. (1997). A spatial extension of CIELAB for digital color-image reproduction. *Journal of the Society for Information Display*, 5:61 – 63.

Zhao, Y., Zhu, C., Yu, L., and Tanimoto, M. (2013). 3D-TV *System with Depth-Image-Based Rendering: Architectures, Techniques and Challenges*, chapter An Overview of 3D-TV System Using Depth-Image-Based Rendering, pages 3 – 35. Springer-Verlag New York.

Zhifeng, S., Jian, H., Mingzhuang, Y., and Mitsuo, M. (2003). Objective evaluation method for appearance of fabric wrinkling by the ccd laser light measuring system. *Sen'i Gakkaishi*, 59:401 – 406.

Zhou, W., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600 – 612.

Zhou, W., Bovik, A., Sheikh, H., and Simoncelli, E. (2014). The ssim index for image quality assessment. `https://ece.uwaterloo.ca/~z70wang/research/ssim/`.

Zuffi, S., Brambilla, C., Beretta, G., and Scala, P. (2007). Human computer interaction: legibility and contrast. In *Proc. of the International Conference on Image Analysis and Processing*, pages 241 – 246.